Foveal Machine Vision Systems

Cesar Bandera

A dissertation submitted to the

Department of Electrical and Computer Engineering

in partial fulfillment of the degree of

Doctor of Philosophy

August 1990

(1)

Foveal Machine Vision Systems

Cesar Bandera

A dissertation submitted to the

Department of Electrical and Computer Engineering

in partial fulfillment of the degree of

Doctor of Philosophy

August 1990

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS N/A
2a. SECURITY CLASSIFICATION AUTHORITY N/A	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for Public Release Distribution Unlimited
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE N/A	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) 605-9160001	5. MONITORING ORGANIZATION REPORT NUMBER(S) N/A
6a. NAME OF PERFORMING ORGANIZATION Amherst Systems, Inc. 6b. OFFICE SYMBOL (If applicable) N/A	7a. NAME OF MONITORING ORGANIZATION U.S. Army Strategic Defense Command
6c. ADDRESS (City, State, and ZIP Code) 30 Wilson Road Buffalo, New York 14221	7b. ADDRESS (City, State, and ZIP Code) Directorate CSSD-H-V Huntsville, AL 35807-3801
8a. NAME OF FUNDING / SPONSORING ORGANIZATION USAMICOM 8b. OFFICE SYMBOL (If applicable) AMAMI-RM-PA-AA-CS	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER DASG60-89-C-0075
8c. ADDRESS (City, State, and ZIP Code) Finance and Accounting Office, Bldg. 8027 Redstone Arsenal, AL 35898-5092	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO / WORK UNIT ACCESSION NO.
11. TITLE (Include Security Classification) Foveal Machine Vision Systems Dissertation	
12. PERSONAL AUTHOR(S) Cesar Bandera	
13a. TYPE OF REPORT Dissertation 13b. TIME COVERED FROM Feb 1990 TO Aug 1990	14. DATE OF REPORT (Year, Month, Day) 1990 August 1 15. PAGE COUNT 292
16. SUPPLEMENTARY NOTATION This dissertation is to replace in full Foveal Machine Vision Systems Final Report dated 1990 February 1.	
17. COSATI CODES FIELD / GROUP / SUB-GROUP	18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number) Image Processing Acquisition Vision Identification Focal Plane Tracking Infrared Imaging Adaptive
19. ABSTRACT (Continue on reverse if necessary and identify by block number) See page(i) of dissertation.	
20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Cesar Bandera	22b. TELEPHONE (Include Area Code) (716) 631-0181 22c. OFFICE SYMBOL

DD Form 1473, JUN 86 *Previous editions are obsolete.*

This work is dedicated to my wife Elisa Victoria.

Abstract

Hierarchical processing is receiving increased attention as a powerful technique for image processing and scene understanding in active vision. Machine vision systems have been implemented using hierarchical processing reminiscent of biological vision with very good results. However, these systems typically employ uniform sampling, which is markedly different from the dynamic spatiotemporal allocation of resolution resources in space variant (foveal) vertebrate vision. Such resource allocation maximizes the relevance of measured scene information to the task at hand and permits more efficient use of the system signal and computational bandwidth. The few space variant sampling strategies proposed in the literature do not lend themselves to hierarchical image processing.

This work presents a new class of active machine vision systems, called *foveal machine vision systems*, which feature space variant sampling directed by gaze strategies. Two families of space variant sampling geometries are analyzed with spatial resolution decreasing with distance from the optical axis. One family features a linear acuity roll-off, and the other an exponential roll-off. Techniques are presented for the integration of sensor frames into unified static scene perceptions. Foveal systems can use many existing hierarchical processing techniques, in particular image pyramid structures and algorithms. A hierarchical structure called the *foveal polygon* is described. The foveal polygon is the subset of an image pyramid supported by foveal sensor frame. Top-down (coarse-to-fine) algorithms processing polygon data serve as drivers for gaze control. Additional gaze control strategies are presented for general learning and surveying (minimization of hypothesis entropy), and feature interrogation (hypothesis likelihood maximization).

Foveal sampling can provide several orders of magnitude reduction in frame size compared to uniform sampling with the same field-of-view and maximum resolution. These savings increase with field-of-view and resolution. The hierarchical data structures generated from the frames are likewise reduced. Analytical expressions are derived for system behavior in the course of localizing unresolved targets. Several redirections of the sensor gaze angle (foveations) are required to accomplish the task, and static savings in frame data produce equally significant savings in overall system bandwidth. Among the two foveal geometries considered, the exponential geometry processes the least data to accomplish the task. The linear foveal geometry exhibits a localization error profile over time similar to that of human saccades performing the same task. Simulations are conducted with resolved objects and hierarchical models. Foveal systems can offer several orders of magnitude savings in computations and data structure size in these more complex settings.

Acknowledgments

The author wishes to express his sincerest appreciation to Dr. Peter D. Scott, Associate Professor in the Department of Electrical and Computer Engineering and in the School of Medicine, Department of Biophysical Sciences, of the State University of New York at Buffalo, for his guidance throughout this dissertation. In addition, the author thanks Joseph V. Fritz for his continuous assistance and encouragement, and Dr. Russ Miller, Associate Professor in the Department of Computer Science of the State University of New York at Buffalo, for his valuable comments. The author also extends his appreciation to the management and staff of Amherst Systems, Inc., for their support of this effort. This work has been supported in part by SDIO contract DASG0-89-C-0075.

Table of Contents

List of Figures

List of Tables

Chapter 1

Problem Statement

1.1 Problem Statement

Active vision refers to the processes of visual sensing and assimilation whereby the platform to which the vision system pertains can reactively maneuver to alter its perception of the scene [Bajcsy88]. Examples of systems performing active vision include a camera mounted on a platform which can move to better resolve features of interest that are partially obscured or far away, or which may have some robotic capability to manipulate or sense objects with an end-effector, again with the goal to better resolve the object [Bajcsy86]. Of course, active vision is performed by all advanced biological systems, including man. Examples of systems not performing active vision include off-line image processing workstations, stand-off sensors, or reconnaissance vehicles following a predetermined path; in the latter case, the platform is moving but not reactively in a context sensitive or perception driven fashion. This work deals with active vision systems where the optical axis and field-of-view of the sensor is steered in a closed loop context sensitive fashion (gazing) within the field-of-regard (the region of space within which system perception is confined).

For many tasks occurring in active vision applications and typical scenes from nature, features that must be resolved in order to accomplish the task are localized in the scene. This localization of "relevance" may appear not only within the field-of-regard of the task, but also within the field-of-view of the sensor. In such cases, uniformly sampling within the field-of-view seems intuitively inappropriate; regions within the field-of-view with little or no relevance to the task are sampled at the same resolution as key features, occupying valuable data storage and signal processing resources. When key features are small with respect to the field-of-view, irrelevant data may dominate these resources. In general, resources are used more efficiently when local system resolution matches the bandwidth of relevant scene features (i.e., neither oversamples nor undersamples a relevant

feature), where the definition of relevance is highly task dependent. This implies that irrelevant features could be subsampled after antialiasing.

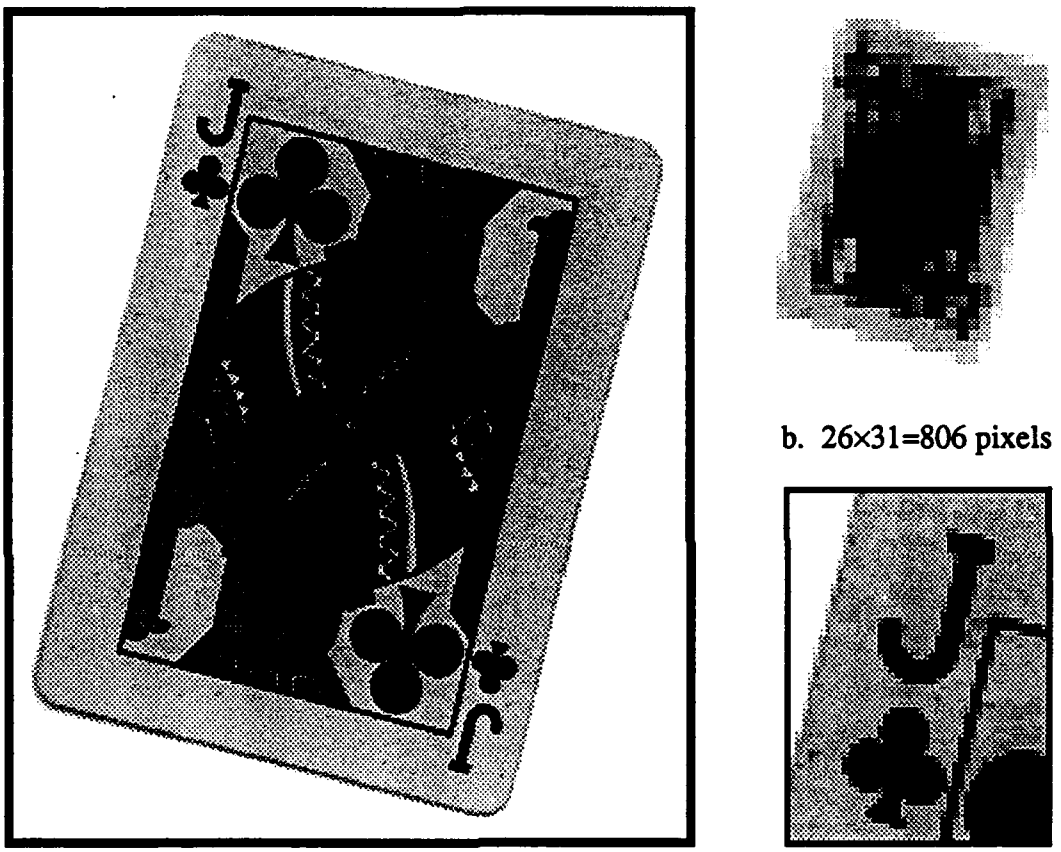
There are two ways by which resolution may be dispersed over a scene which is to be studied. The spatial resolution may vary over its field-of-view, yielding images with finer detail in some image regions than others, or the field-of-view may be allocated temporally, yielding longer gaze times and shorter revisit times to some regions of the scene than others. For static scenes in the presence of noise and sensor imperfections, the effective accuracy of measurements at any point in the scene (i.e., localized resolution) improves monotonically with each of these factors. This relationship may vary for other scenes, but effective localized resolution remains determined significantly by the spatio-temporal allocation of resolution resources.

Current multiresolution systems employ uniform resolution sensors and generate a hierarchical multiresolution feature set. Consider, for example, the hierarchical model of an object as shown in Figure 1.1-1. The object is represented by large features requiring little resolution to measure, such as overall aspect ratio, and small cues (e.g., the card number and type on the top left corner) requiring greater but localized resolution. In this case, sampling the entire object at the high resolution necessary to analyze the small cue (Figure 1.1-1a) produces two orders of magnitude more data than that necessary to perform area measurements (Figure 1.1-1b), and 40 times more data than locally sampling the small cue (Figure 1.1-1c).

In the event that measurements are significantly corrupted by time varying sources, such as sensor noise, glint, and atmospheric aberration, repeated measurements of relevant points in space must be performed to obtain scene estimates of acceptable error variance. Returning to the example, edge analysis of the card letter is inherently more susceptible to such measurement error than area analysis of the overall card. Consequently, the gaze time necessary for the low spatial resolution registration of the card (Figure 1.1-1b) may be significantly less than that for the high spatial resolution registration of the card letter (Figure 1.1-1c), and the processing of frames registering the former can be reduced.

The gaze control strategy of an active machine vision system determines what scene regions are to be observed, how long the gaze is to be maintained on a feature, and possibly how soon the feature should be revisited. Gaze control by active feedback requires sophisticated mechanical servo loops, whose dynamics must be based on unknown strategies for deciding where to look next and for how long. The study of such

"active perception" systems and strategies, pioneered in the laboratories of Bajcsy, Burt, Ballard and Aloimonos, while of relatively recent origin, is achieving recognition as an important and continuing line of investigation in machine vision and machine perception [Aliom87], [Bajcy88], [Balla87], [Brown89], [Burt88]. Continued progress in this field is vital to the development of the types of vision systems proposed by this work.



a. 262×316=82,792 pixels

b. 26×31=806 pixels

c. 45×60=2700 pixels

Figure 1.1-1. Hierarchical representation of a playing card.

In addition to the objectives of conventional active machine vision, the gaze control strategy of a system with a multiresolution sensor must orient the optical axis such that scene features required by the task to be highly resolved are registered by the sensor's high resolution region within the field-of-view. Likewise, the low resolution regions in the sensor field-of-view should register scene features requiring less resolution, or which appear in the field-of-view but are irrelevant to the task. In this fashion, scene features are

resolved to the degree required by the vision task (including minimum resolving of irrelevant features), conserving data and computational bandwidth [Schri88]. The *vision task* is defined here as the task which the machine vision system must accomplish.

Less progress has been made on space variant systems. This is evident by comparing the wide repertoire of existing and well understood processing tools for uniformly sampled images with the few implemented space variant algorithms. Even hierarchical machine vision systems, while being multiresolution in the general sense, operate at discrete levels of typically uniform resolution [Uhr86]. These systems offer significant performance advantages over traditional image processing at the low and medium level steps of the overall vision process [Rosen88], [Scott90]. Additional significant improvements in performance could be expected if the hierarchical nature of the vision system would commence at the sensor and the raw data format itself. Furthermore, a greater portion of the vision process would benefit from these improvements.

Space variant vision approaches with space variant sampling have been proposed but in a fashion that precludes the straightforward use of current hierarchical processing techniques [Katsi87], [Porat88], [Yeshu89], [Zeevi88]. The benefits of data bandwidth reduction are negated if efficient image processing techniques must be replaced by more complicated operations, such as spatial/spectral warping or complex interpolation of data to generate hierarchical levels. Here, image processing concepts intuitive in the uniform case, such as signal bandwidth and nearest neighbor samples, become unclear.

While space variant active vision is novel in the machine setting, it is realized throughout nature, including by human vision. The human retina has a resolution which is minimum at the periphery of the field-of-view and two orders of magnitude greater at the optical axis [Yarbus67]. This resolution function is fixed. The prevalence of this approach in nature provides a certain degree of justification for the study of its implementation in the machine setting (billions of years of evolution cannot all be wrong).

1.2 Outline of Thesis

This dissertation presents and investigates a new family of graded resolution active machine vision systems. The specific scientific contribution of this work is a collection of multiresolution sensor geometries, gaze control strategies, and multiresolution data processing algorithms and architectures. The objective of this work is not to mimic the vertebrate visual system but to improve the cost effectiveness of active machine vision systems. Nevertheless, several interesting analogies are observed between both types of vision systems, and a number of works in the field of physiology are referenced.

Chapter 2 introduces concepts of graded acuity active vision in both the biological and machine setting. Chapter 3 analyzes the geometric properties of graded acuity sensor frames. Several families of graded resolution sensor geometries are presented. Analytical solutions for system bandwidth are derived for the vision task of finding an unresolved target against a dark background. From these results, significant reductions (two orders of magnitude or more) in data and computational bandwidth are observed with respect to uniform systems. Also, an optimum sensor geometry for the target localization task is determined. Chapter 4 presents techniques for the generation of integrated perceptions of static scenes. These integrated perceptions retain and fuse the information from a sequence of graded resolution sensor frames. Chapter 5 presents gaze control strategies which orient the sensor so as to maximize the collection of relevant information. Chapter 6 presents hierarchical data structures, algorithms, and computer architectures for the implementation of gaze control algorithms and the processing of images (sensor frames and integrated perceptions) from graded acuity machine vision systems. Conclusions and topics for further research are given in Chapter 7.

Chapter 2

Introduction to Foveal Machine Vision

2.1 Overview of Foveal Vision Approach

In this work, we investigate the properties of a new class of machine vision systems for active vision called *foveal machine vision systems*. The term "*foveal*" originates from the fact that the system adopts from biological vision the attribute of variable resolution sampling and a high resolution region in the sensor field-of-view called the *fovea*. As with conventional active vision systems, the field-of-regard for the task is interrogated by steering the optical axis and field-of-view of the sensor (gazing) in a closed loop context sensitive fashion. Foveal systems distinguish themselves from uniform sampling resolution systems by the fact that through intelligent gazing, the system allocates resolution and processing resources to those features of interest within the field-of-view, and minimizes the expenditure of these resources on regions irrelevant to the task being performed.

The graded resolution of foveal systems allows the field-of-view to be widened at low resolution while incurring negligible system cost. Thus, the hard spatial resolution versus temporal resolution (field-of-view) trade-off imposed on uniform sampling systems is relaxed. High spatial resolution is maintained locally at the fovea, while a wide field-of-view minimizes revisit times and improves system performance in scenes with uncooperative moving objects.

By dynamically allocating visual resolution (spatial and temporal) to objects and object features of relevance to the task, sensor and processing resources are used more efficiently. This, in turn, can support a decrease of several orders of magnitude in the amount of data processed in the execution and completion of vision tasks. Tasks may be completed faster, and complex image processing algorithms which require additional processing can be implemented while maintaining real-time performance. If the size and cost of the vision system takes precedence over speed, then the foveal system may be implemented with much less processing hardware than a uniform resolution

implementation, while maintaining comparable task execution speed. Equivalently, the processing resources of a uniform resolution system may be used in a foveal system with greater sensor field-of-view and localized resolution, thus keeping bandwidth constant but increasing system performance figures of merit such as detection range and revisit time.

2.2 Attributes of Foveal Vision

Foveal machine vision systems can be implemented with the same base technologies in current vision systems (e.g., sensor material, computer hardware, image processing algorithms). A unique feature of foveal systems is their novel architecture which adopts a number of attributes from biological foveal vision. This is described in the following sections. Specifically, foveal vision requires a tighter coupling of the different levels of vision processing than what appears in current machine vision systems [Rosen88].

2.2.1 Multiresolution Sampling

Variable resolution scene sampling can be achieved in various different ways. A uniform resolution camera with a zoom lens accomplishes this feat while retaining the advantages of uniform sample spacing. On the other end of the complexity spectrum, one can imagine a dynamically reconfigurable focal plane array which changes its sensor element pattern in real-time to any of a nearly unlimited number of nonuniform geometries.

A critical drawback of the first implementation is that when interrogating relevant features with high resolution, the system suffers from acute "tunnel vision," as resolution and field-of-view are still played against each other. In dynamic scenes where the vision system must always be on the alert for approaching objects (e.g., navigation, detection of incoming objects, defensive "prey mode") or when an object under interrogation can at any instant move rapidly or deviate from an estimated trajectory (e.g., object tracking, pursuit, offensive "predator mode"), a wide field-of-view simultaneously maintained with localized high resolution is critical to the successful execution of the vision task (or survival of the organism).

The dynamically reconfigurable focal plane array has two significant drawbacks. First, and most obvious, is the complexity of such a sensor. This sensor could be implemented with dynamically reconfigurable mirrors, a technology used to correct for atmospheric aberration in astronomy and defense applications. Another implementation is to perform localized decimation or averaging within a frame generated by a uniform sensor which simultaneously meets system field-of-view and resolution requirements, although to a certain extent this is self defeating. Either way, the control overhead for all the parameters of such a device would be extensive.

Second, while a perfect match between localized feature bandwidth and sensor resolution is conceptually possible, extensive knowledge of the actual feature would be required. Feature identity is not known deterministically by the vision system but is instead hypothesized (otherwise, there would be no need to interrogate it). One is thus faced with the law of diminishing returns, whereby the extensive control overhead of such a sensor may outweigh the cost of slightly suboptimal resolution matching and bandwidth utilization, especially when taking into account the probability of feature misclassification.

Nature has opted for a compromise solution. The sensor (retina) has a non-reconfigurable structure which offers a required range in resolution. The distribution of resolution within the field-of-view can be designed so as to optimize the performance of the vision system in the execution of particular tasks. Such a sensor can also be implemented in a machine setting in any of several different ways as discussed in Chapter 7. Because of sensor feasibility, efficient bandwidth utilization, and no control overhead other than gaze control, this type of sensor seems to be the most appropriate engineering solution for the foveal machine vision system.

A polar sampling focal plane array has been developed with an order of magnitude higher resolution at the center than at the periphery [Kreid90], [Vander89]. This is the only machine sensor known to exist by the author with a multiresolution monolithic design, justifying to an extent the selection of a fixed geometry foveal sensor.

Of critical importance in the implementation of the fixed multiresolution sensor is the selection of the resolution roll-off, that is, the rate in which resolution decreases with distance from the optical axis. This is as important to the dynamic performance of the active vision system as the more conventional parameters of field-of-view and maximum resolution. The resolution roll-off from the fovea is determined by the localization of the

bandwidth of relevant scene features. The size of the fovea is determined by the expected scale of the more relevant features. The sensor should be able to resolve this feature uniformly with its highest resolution. Examples are given later in this chapter illustrating this optimization in nature.

2.2.2 Gaze Control

Conventional active vision performs spatio-temporal allocation of resolution within the field-of-regard, but not within the field-of-view (an obvious consequence of uniform sampling by the sensor within its field-of-view). The foveal system performs a more refined resource allocation whereby sensor resolution is matched with the resolution of the relevant localized scene information.

As with human vision, it may not be possible to resolve a relevant scene feature entirely within a single sensor frame with sufficient resolution (where the task defines sufficient). This occurs when the feature is scaled such that it is larger than the subset of the sensor field-of-view with the specified resolution or greater. In effect, scene information exceeds the spatial resolution of the sensor. Consequently, the sensor axis must be scanned over the feature so as to collect the necessary data.

When several detected features require further interrogation at higher spatial or temporal resolution (*cues*), the sensor may be able to satisfy the information requirements of the vision system on several of these simultaneously with a single redirection of the sensor's optical axis, or *foveation*. In this case, the gaze control problem can be posed as one of statistical control and queuing techniques with feedback from the higher vision processing levels to the low pixel level. In conventional uniform sampling vision, this feedback controls the resampling of an image frame [Weems86]. In foveal systems, the feedback spans the entire vision system all the way to sensor orientation.

The foveal system orients its sensor(s) so as to maximize the acquisition of relevant information. A reasonable question to ask is, "How does the system know where the information exists, and with what resolution must it be registered?" The solution lies in the gaze control strategy of the system, and at the heart of the strategies considered in this work is the collection of object models.

2.2.3 Hierarchical Models

As with any application of scene understanding, models of objects relevant to the task are a prerequisite. In foveal vision, the model is hierarchical such that the object is described in terms of strong (easily perceived) features at different resolutions (e.g., for the playing card, the features could be aspect ratio and a letter at the upper left hand corner).

The dynamic aspect of gaze control begins when (1) a scene feature is perceived which matches with a known feature of the object model, and (2) a hypothesis can be made on the identity of the feature. The model itself then becomes the roadmap for the search for additional information on that object, and the system looks at locations where the model predicts additional corroborative features. The hierarchical model also gives the resolution necessary to properly register the individual features, so the sensor resolution can be matched with the bandwidth of the feature or groups of features under interrogation.

2.2.4 Hierarchical Integrated Perception

Human perception is very stable in comparison with the constant refixation of gaze in three dimensions (eye movement and refocusing). Likewise, machine perception should not be defined in terms of local sensor coordinates, but in terms of more stable coordinates like that of the vision system itself, the general environment (inertial reference frame), or key objects of relevance to the vision task.

Foveal systems require multiple sensor frames or views at different gaze angles to accomplish any vision task of reasonable complexity. The integrated perception retains the salient features detected in the field-of-regard. Under successful gaze control, these features are efficiently represented by the sensor frames. In static scenes, a low level perception may be generated by fusing the frames into a multiresolution view of the field-of-regard which resolves the salient features with higher acuity. In dynamic scenes, objects are moving and frames may not correlate, so the integrated perception must be formed at a higher level, such as integrating feature or symbolic data resulting from the processing of individual frames or frame sequences.

2.3 Reference to Biological Vision

Biological vision is considered in this effort for several reasons:

1. Biological vision performance exceeds that of man-made systems in most respects.
2. There are fundamental attributes of biological vision systems which have not been applied to man-made systems, and which are partially responsible for the formers' superior performance.
3. Biological vision has been optimized through natural evolution for the efficient accomplishment of dynamic imagery tasks.
4. Psychophysical research provides information on biological vision, which serves as reference data to supplement current digital signal processing theory.

Current digital processor technology operates at nanosecond response time. The processing components comprising biological vision systems operate at response times on the order of ten milliseconds, yet dramatically outperform man-made systems in most applications. Even the conceptually simple process of colorizing frames of vintage black-and-white movies, which is only recently feasible with state-of-the-art technology and which is analogous to drawing in a child's coloring book, requires extensive human intervention [Fisch87]. This implies that the format in which image data is represented and processed in biological systems is greatly superior to that of man-made systems.

Human visual resolution is sharpest at the center of the field-of-view along the eye's optical axis, where the fovea is situated. The cones are the photosensors in the human retina responsible for high resolution. The fovea is characterized by a high concentration of cones relative to rods; a 1° area of the fovea consists exclusively of cones. The relative concentration of cones to rods decreases with angle from the optical axis. At the parafovea (8.6°), there are less cones than rods, and the relative concentration drops sharply in the perifovea region extending from 8.6° to 19°. Beyond the perifovea is the peripheral retina which has two orders of magnitude less acuity than the fovea (Figure 2.3-1).

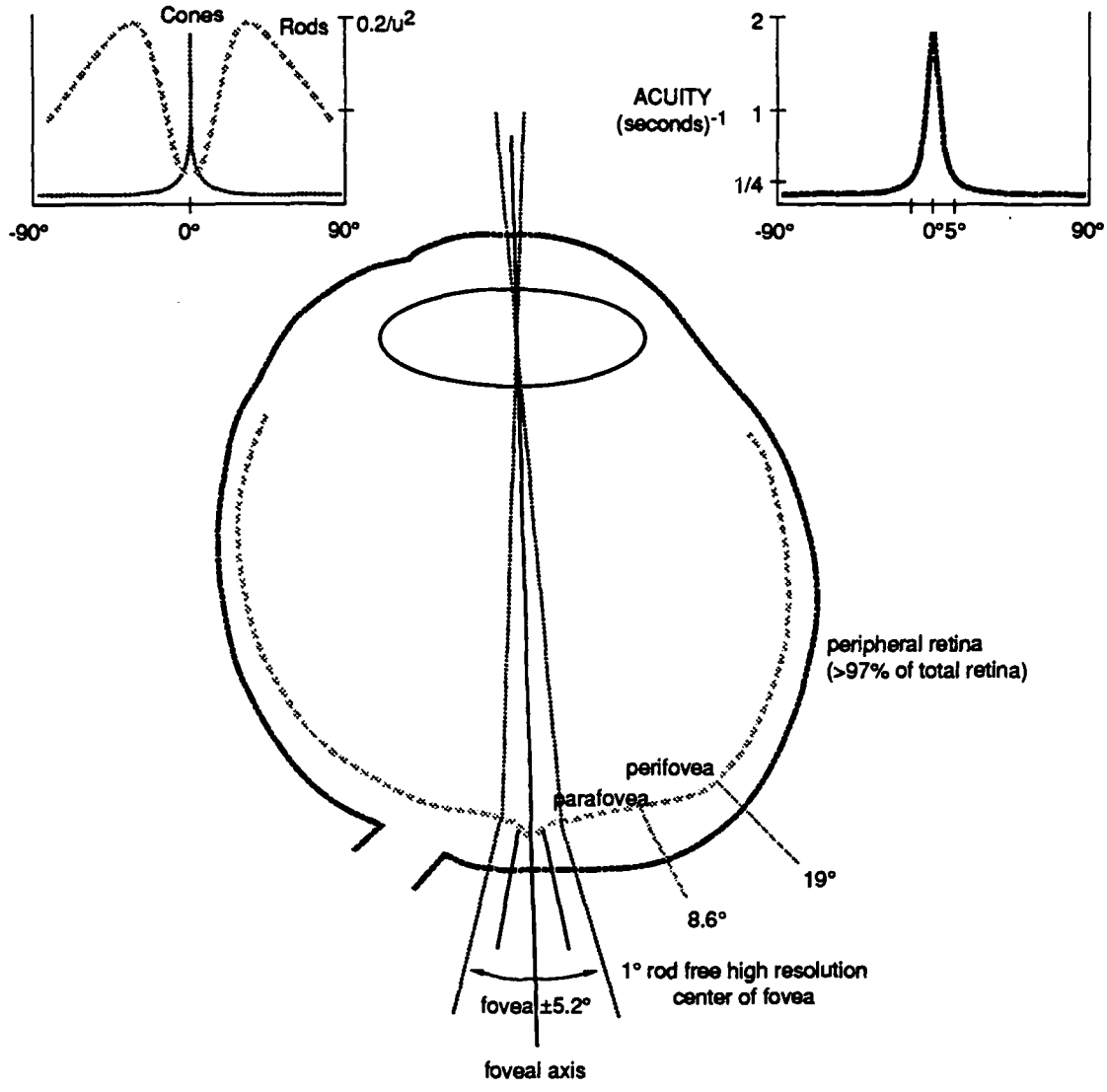


Figure 2.3-1. Human visual acuity.

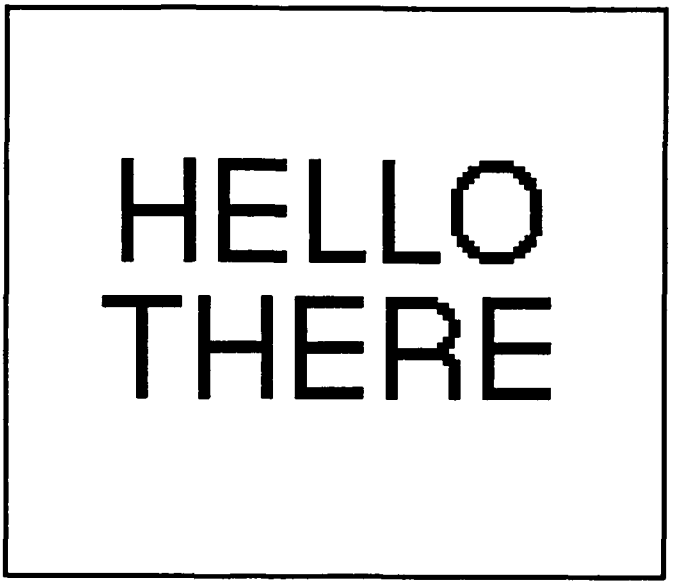
Acuity is expressed as the reciprocal of the angle, subtended at the eye, of the smallest perceptible scene feature; the greater the acuity value, the greater the resolution in that region of the field-of-view. Visual acuity, motion detection, and other psychophysical measurements are no more than inversely proportional to the angle from the optical axis [Burbec87]. This peaked acuity is why, for example, the reader of this document must scan his or her gaze across the page to read the text, even though the entire page is within the field-of-view at all times; the fovea is required to resolve the letters of text, which are unrecognizable blurs to the peripheral vision. As image data is received by the visual and association cortex, it is integrated into a single perception of the page.

It should be noted that biological vision benefits from many other architectural attributes, such as the "monolithic" processing at the retina including difference-of-gaussian spatial filtering, temporal filtering, and aperture control [Marr76]. Some of these features have been implemented in a machine setting [Lubkin90].

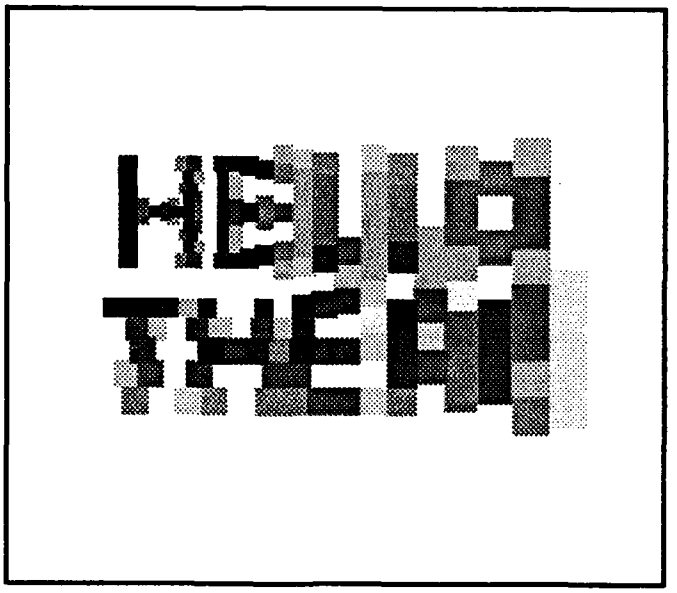
Figure 2.3-2 illustrates the concept of variable resolution images. The image presented in Figure 2.3-2b was generated by a simulated sensor with variable resolution similar to that of the human retina as it gazed over the center of the first "E" in the scene presented in Figure 2.3-2a. The reader may argue that he or she can see the top illustration much clearer than the sensor. The bottom illustration does indeed represent the information provided by the retina. However, what the reader *perceives* is the integration of detailed information from up to a dozen image "frames" taken as the eye continuously performs minute reorientations of the optical axis called microsaccades. Overall data bandwidth savings are obtained even after generating the frame sequence because the space variable resolution reduces the signal bandwidth from the eye by a factor of approximately 1:16,000 below that if the fovea acuity extended throughout the field-of-view [Yeshu89].

The shape of the biological acuity profile is typically fixed for a given organism, discounting the secondary effects from ocular aperture variation, spectral aberration from focusing, and eye muscle induced structure distortions. Through natural evolution, the profiles have attained a shape which optimizes the performance of the vision system when conducting tasks critical to the life of the organism [Levin85]. For example, small mammal prey such as rabbits have a fovea that is wide and narrow to help detect predators approaching on the ground or the shadows of airborne predators. Predators, including man, have an acuity profile which is circularly symmetric to support the chase of prey. Several organisms even have two foveae (bimodal acuity profile); the eagle has one fovea oriented forward for flying and another oriented downward for detection of prey, and the fish *Anableps microlepis* has one fovea oriented upward and another downward for detecting prey above the water surface (insects) and below (small fish).

The size of the fovea in animals is also optimized to the more critical tasks performed by the organism. The human fovea, for example, resolves a small bite-sized food item at arm's length (humans and monkeys distinguish themselves from other vertebrates by bringing food to their mouths as opposed to the converse). The complete two-dimensional acuity profile presents to the systems engineer a rich degree of freedom in the design of foveal sensors permitting them to be architecturally tuned to the critical vision tasks.



a. Scene (350×300=105,000 samples).



b. Image generated from a sensor mimicking the human retina (647 samples).

Figure 2.3-2. Example of space variable resolution sampling.

Eye movements, or *foveations*, in primates have been classified into two gaze control strategies: saccadic and foveal pursuit eye movements [Eckmil83]. These strategies refer to eye gimballing only, and do not involve head movement. Saccadic movement refers to the abrupt repositioning of the sensor axis from some initial position to another position within the field-of-view (±20°) of the first gaze. In this fashion, the vision system interrogates with the fovea cues uncovered by low acuity peripheral vision. During

saccadic eye movement, it is believed that sensor output is "disconnected", primarily because the movement is so fast (over 600° per second) that the vision system cannot correlate the data into a stabilized integrated perception.

Foveal pursuit eye movement refers to the tracking with the fovea of a target moving in the field-of-view. The human system is capable of tracking objects with velocities exceeding ±50° per second and acceleration exceeding 250° per second squared. Motion estimation is performed because there is no permanent tracking lag. If the object outmaneuvers the eye, by exceeding it dynamics and/or changing abruptly from the estimated path, the eye re-establishes track with a saccadic eye movement.

Current active vision systems perform sensor movement similar (in a coarse way) to foveal pursuit. Here, the objective is to maintain the optical axis at or near the centroid of the moving object. However, there is no corresponding strategy in the uniform case to saccadic eye movement. The uniform system may try to get an entire object within the field-of-view, but this is categorically different from saccadic movement, which is the allocation of spatial resolution within the field-of-view, not just within the field-of-regard. This work addresses the implementation of saccadic gaze control strategies for the machine vision setting.

Chapter 3

Foveal Geometries and Saccadic Performance

3.1 Introduction

This chapter presents and analyzes two different foveal acuity profiles, called the linear and exponential patterns, which characterize very different families of sensor geometries. These patterns differ, among other respects, in the degree of resolution roll-off with distance from the fovea. The localized acuities present in the exponential pattern are related by powers of two, making it more analytically tractable and resulting in a frame data format which lends itself to hierarchical processing.

Sensor element count and frame data size are derived for each pattern. Very significant reductions in the amount of frame data are obtained by reducing acuity in the periphery of the field-of-view. The dynamic attributes of foveal system operation are also investigated, such as the number of sensor axis relocations, frame data bandwidth, and the total amount of data processed in the execution and completion of an objective, or vision task. Dynamic attributes can only be estimated in the context of a vision task and stochastic model of the scene. The task of localizing a bright unresolved stationary target against a dark background is employed as the reference set of conditions. A saccadic gaze control strategy is implemented; this strategy is unique to foveal systems. Analytical expressions are derived for task completion times, processed data set size, and computational statistics. These expressions are confirmed through simulation.

The total amount of data generated multiplied by the number of time steps involved in the processing of the data, with each processor instruction occupying one time step, is used as a complexity measure. The inverse of the complexity measure is adopted as a figure of merit. The complexity measure of a conventional uniprocessor uniform acuity vision system performing the task of unresolved target localization within an N pixel by N pixel scene is $O[N^4]$. The expected complexity measure for uniprocessor foveal systems are $O[N^2]$ if the linear pattern is used, and $O[(\log_2 N)^3]$ if the exponential pattern is used.

The worst case complexity measures for the linear and exponential foveal systems are $O[N^2\log_2(\log_2 N)]$ and $O[(\log_2 N)^3]$, respectively (the expected and worst case exponential measures differ by a small second order term).

It is shown that foveal systems can offer a significant savings in data and computational bandwidth when compared to uniform acuity vision systems. Furthermore, these savings increase with system field-of-view and maximum resolution, making foveal systems particularly attractive to demanding applications where conventional solutions require multimillion pixel focal plane arrays. Again, localization of relevant information in the scene is necessary for good foveal system performance.

3.2 Attributes of Foveal Focal Plane Arrays

In this work, the *focal plane array* (FPA) is used as a reference sensor implementation because there is a direct correspondence between sensing element geometry and acuity. Specifically, the far field acuity of an FPA is inversely proportional to the linear size of the sensor elements.[1] A foveal FPA can thus be constructed by using variable sized sensor elements as opposed to the uniformly sized elements of conventional FPAs. The acuity profile of the foveal FPA is shaped by sizing the individual sensor elements to obtain the desired acuity at the corresponding region in the field-of-view.[2] This work will consider foveal FPAs with sensor element scaling which is is monotone nondecreasing with distance from the optical axis.

Quantum FPA efficiency is proportional to the ratio of active (photosensitive) surface area to overall FPA surface area. In order for this ratio to be 1, the elements must tile perfectly such that there is no space between them. This imposes a fundamental constraint on the FPA sensor element geometry. Actual implementations of FPAs cannot achieve 100% utilization of the surface area as active sensor area because the boundaries isolating the sensor elements themselves occupy space. Current manufacturing techniques

---

[1] $\sin(\alpha)\equiv\alpha$ for small α.

[2] The relative scaling of the independent and dependent axes of the acuity profile depends on the magnification factor of the front end optics; doubling the magnification doubles the acuity and scales by 50% the field-of-view.

have surpassed 85% surface area utilization for infrared sensors [Bothw87], and approach 100% for visible wavelength sensors [Photo89].

Perfect tiling avoids the aliasing that would occur if the elements were small and distributed with interelement spacing on the order of element size. If an acuity profile varying by a factor of 10 between the fovea and the periphery were implemented by elements of the same size but spaced non-uniformly, the sensor would suffer from extreme aliasing at one tenth the bandwidth of the fovea. Perfect tiling minimizes aliasing by lowpass filtering (window averaging) the scene signal in the analog domain.

Whereas photodetectors with perfectly tiled variable sized coverage are prerequisites for foveal sensors, a third property, integer scaling of sensor elements, is very desirable. Integer scaling simplifies the construction of monolithic VLSI sensors. VLSI design rules begin with the definition of a length unit λ against which all pattern features are measured [Mead90]. Integer scaling preserves this geometric normalization. Integer scaling also simplifies the process of forming an integrated perception (discussed in Chapter 4) and the development of closed expressions for system performance. Integer scaling also simplifies the generation of new patterns matching desired acuity and spectral profiles [Jarske88]. Integer scaling disqualifies geometries obtained through "rubber sheet" distortion of uniform grids (Figure 3.2-1). Polar sampling patterns (Figure 3.2-2) can satisfy this constraint with the proper selection of ring radii. An additional constraint on sensor element geometry, which furthers the benefits of integer scaling, is the translation invariance of element shape (not size).[3]

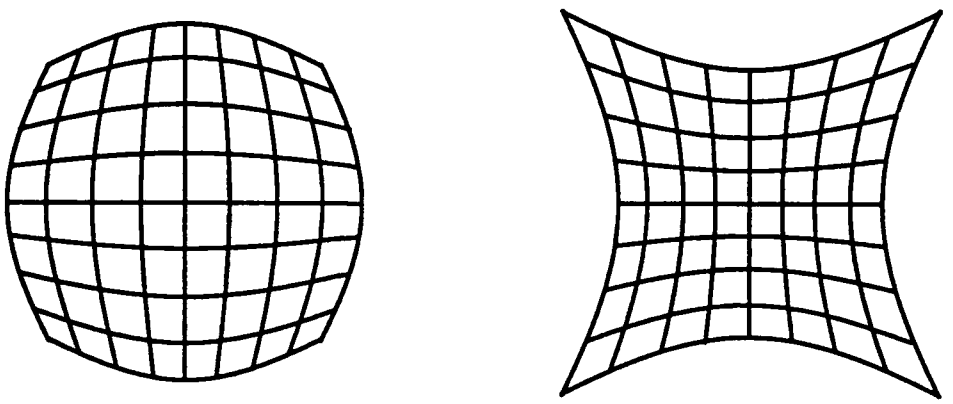


Figure 3.2-1. Rubber sheet distortion and non-integer scaled elements.

[3] At first glance, it would seem that polar sampling patterns do not feature translation invariant element shapes. However, if translation (and orientation) is defined in r-θ space, then translation invariance is preserved.

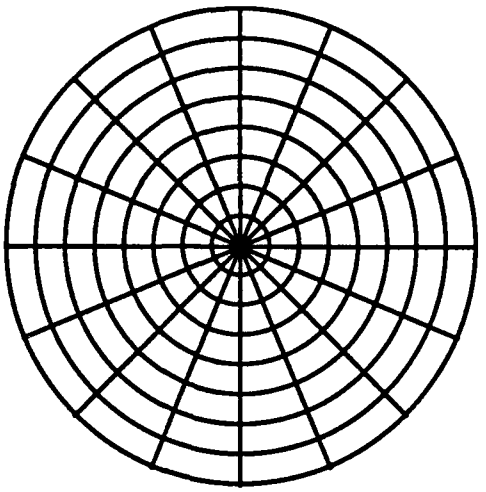
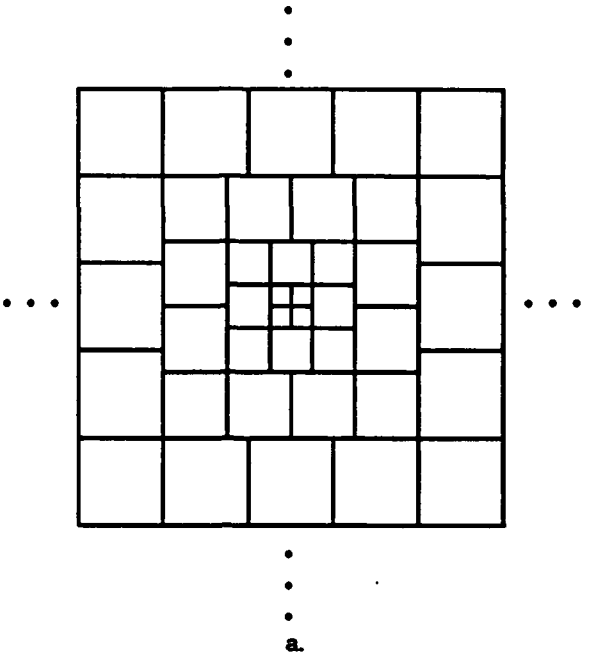


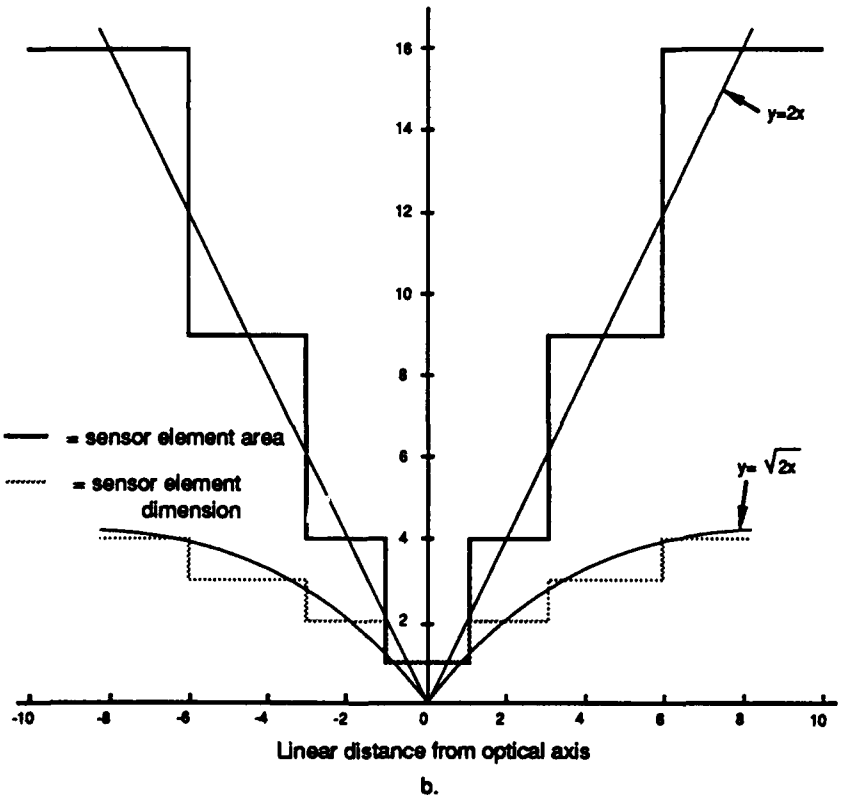
Figure 3.2-2. Polar sampling pattern.

Figures 3.2-3 and 3.2-4 present two different foveal FPA sensor element geometries and their acuity profile along the field-of-view horizon. Both consist of concentric square rings of square sensor elements; the elements vary in size throughout the FPA, but all elements within a ring are the same size. All patterns offer tiled geometry (no blind spots), integer scaling of the tiles, translation invariance of element shape (all elements are square), and a directed allocation of acuity. The geometry in Figure 3.2-3 features sensor element area which grows linearly with distance from the optical axis crossing the pattern center. The linear dimensions of the elements also increase linearly with ring index (the linear dimensions of an element in the i'th ring from the center is proportional to i). This pattern of FPA sensor elements is named the *linear foveal pattern.* It features an inverse radical acuity profile, whereby the reciprocal of the acuity (the element linear dimension) is proportional to the square root of twice the distance from the focal axis. The fovea of the linear pattern consists of a small two by two grid of high resolution sensor elements.

The geometry in Figure 3.2-4 features sensor element area which grows exponentially by a power of two with distance from the optical axis. The linear dimensions of the elements also increase exponentially with ring index (the linear dimensions of an element in the i'th ring from the center is proportional to 2^i). This pattern of FPA sensor elements is named the *exponential foveal pattern.* It features an inverse linear acuity profile, whereby the reciprocal of the acuity is proportional to the distance from the focal axis. The fovea of the exponential pattern consists of a small four by four grid of high resolution sensor elements.

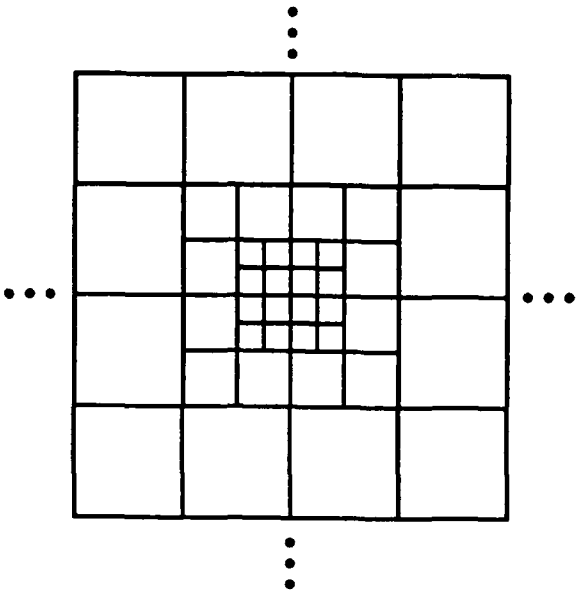


a.

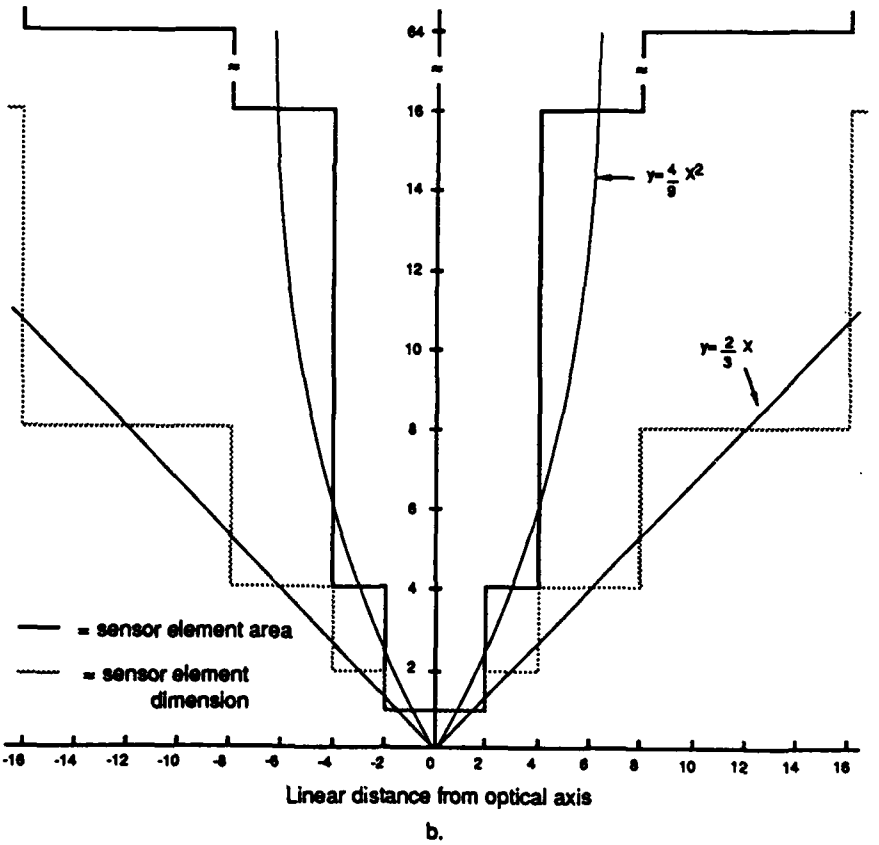


b.

Figure 3.2-3. Linear Foveal Pattern. The geometry is illustrated in (a), and the sensor element area and size are given in (b) with respect to L_{oo} distance from the pattern center. The smooth functions in (b) illustrate the tendency of the element geometry.



a.



b.

Figure 3.2-4. Exponential foveal pattern. The geometry is illustrated in (a), and the sensor element area and size are given in (b) with respect to L_{∞} distance from the pattern center. The smooth functions in (b) illustrate the tendency of the element geometry.

There are many more patterns which offer tiled geometry with a single tile shape, integer element scaling, and a directed allocation of sampling resources. For example, selected sensor element in the linear and exponential patterns may be subdivided to provide a wider fovea or a different acuity roll-off from the optical axis. Figure 3.2-5 illustrates some additional foveal patterns. The first two patterns in Figure 3.2-5 are the linear and exponential patterns, respectively, with larger foveae obtained through subdivision of the sensor elements of the second ring. Figure 3.2-5c is a tri-symmetric exponential (base 2) pattern. The last two patterns are exponential and linear hexagonal tessellations of triangles, respectively. Subdivision of sensor elements permits the design engineer to tailor the acuity profile while maintaining the necessary properties of foveal geometries.

In both conventional uniresolution and foveal FPAs, each sensor element generates one datum per sensing time. The set of data from all elements for a given sensing time is the raw image frame. The datum composing a foveal FPA image will be referred to as a *rexel*, for *res*olution *el*ement, to differentiate it from a pixel in conventional FPA images. Since the size of a sensor element and its location within the FPA are the same as the size (weight) and location of the corresponding datum in the raw data frame, the terms pixels and rexels will also be used to denote the sensor elements themselves.

This section compares the data generated by a foveal FPA (number of rexels) to that of a conventional FPA (number of pixels) with the same field-of-view and maximum resolution. Unless otherwise specified, it will be assumed that all FPAs have the same front-end optics. Furthermore, the relative scaling of the foveal and homogeneous lattice geometries will be such that the smallest rexel in the foveal geometry has the same dimensions as a pixel in the uniform FPA. Thus, dimensions will be normalized to that of a single square pixel (i.e., all pixels are of size 1×1, and a rexel of size $n \times n$ encompasses the area of n^2 pixels).

Figure 3.2-6 illustrates the rexel data frame obtained by sampling a scene with the linear pattern of Figure 3.2-3, the exponential pattern of Figure 3.2-4, and an exponential pattern uniformly subdivided (and renormalized) by a factor of four (each rexel is subdivided into a cluster of 4×4 rexels). The scene consists of 512×512=262144 pixels, whereas the linear frame contains 1104 rexels, the undivided exponential contains 100 rexels, and the subdivided exponential contains 1216 rexels.

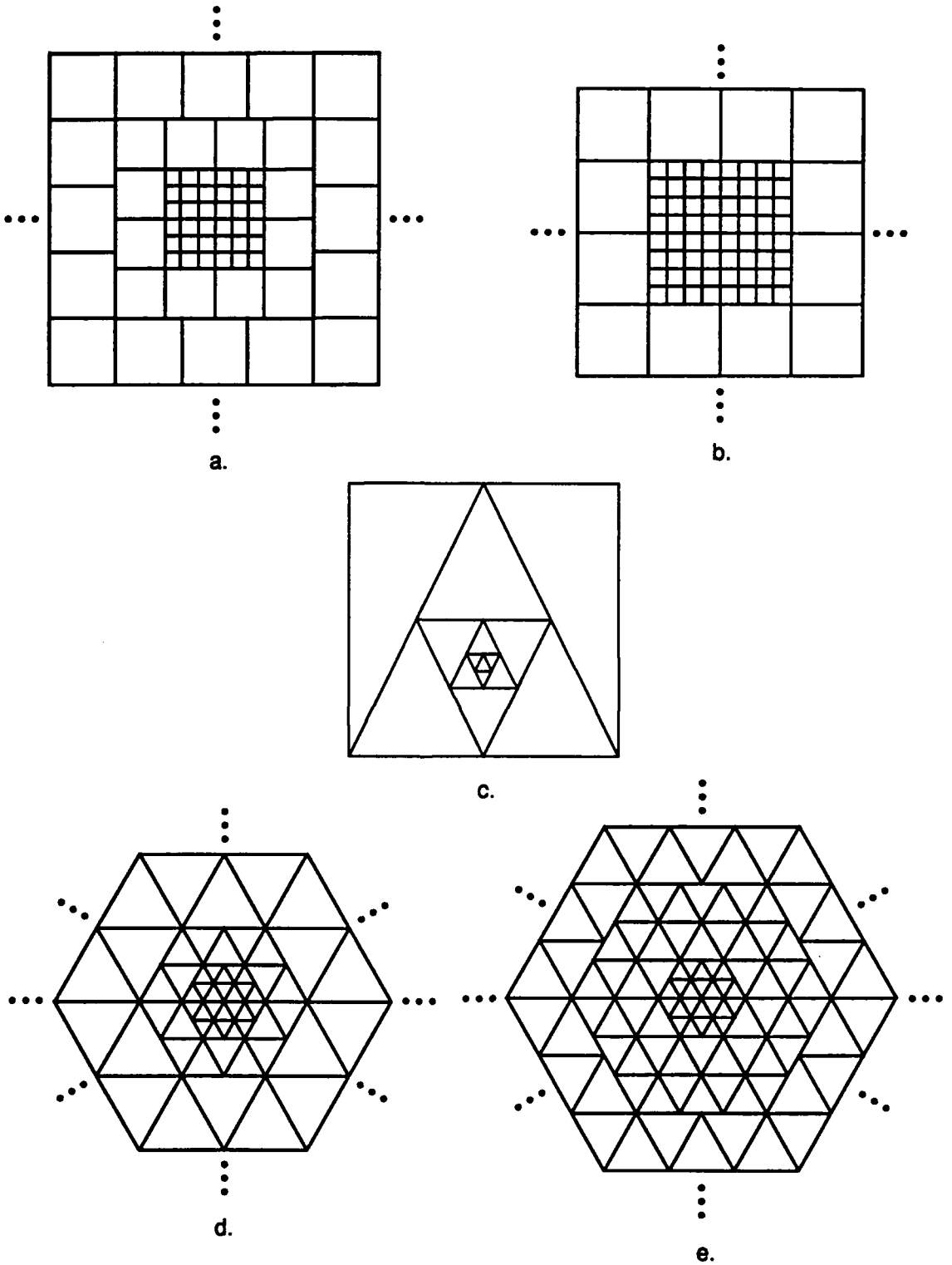
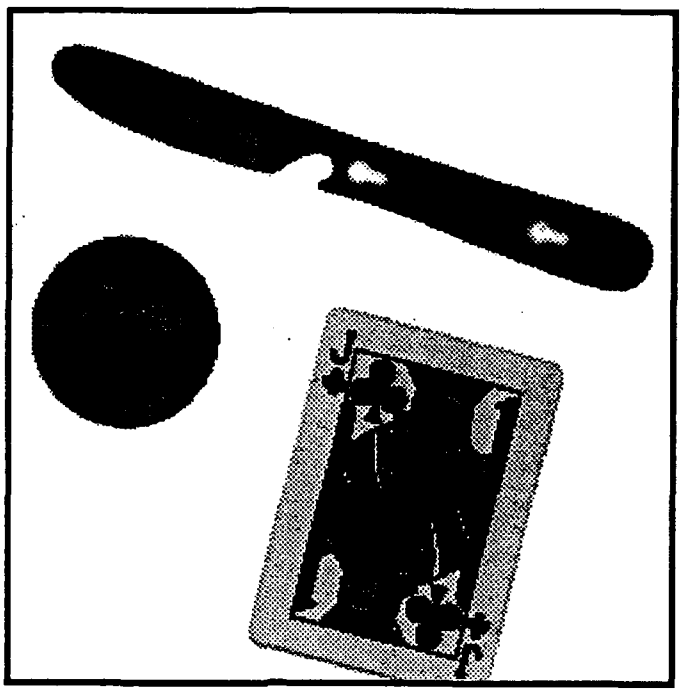
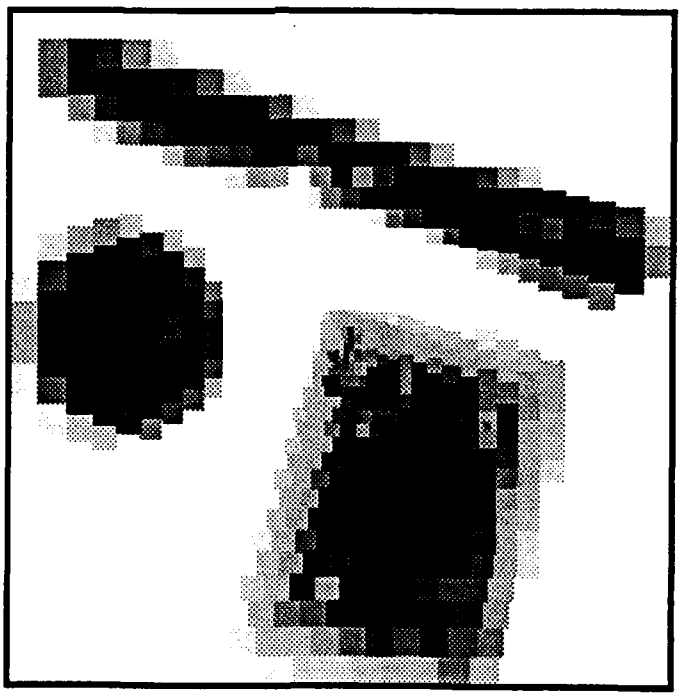


Figure 3.2-5. Additional examples of foveal patterns.

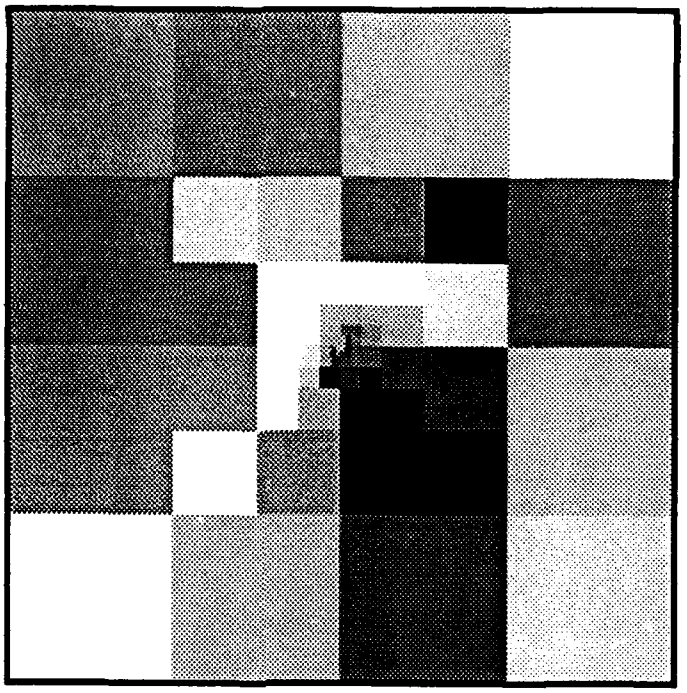


a. Original scene (512×512 pixels).

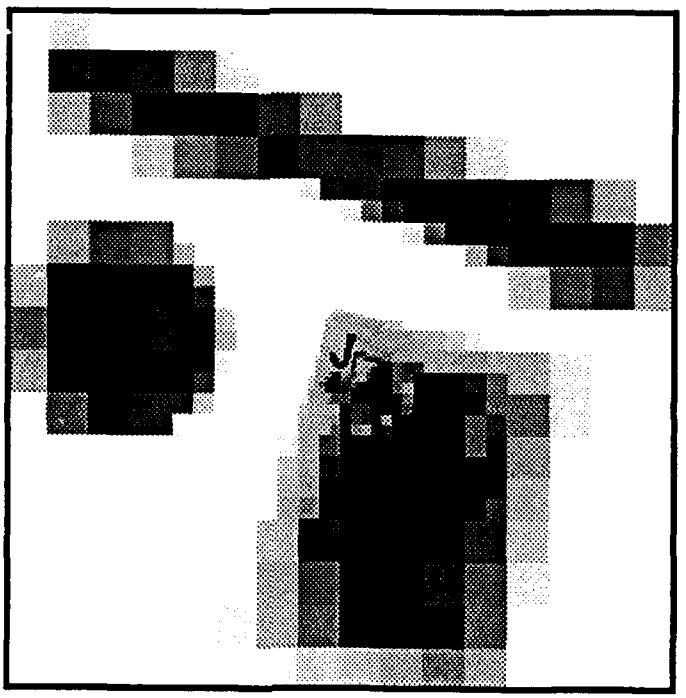


b. Scene sampled with linear foveal geometry.

Figure 3.2-6. Examples of foveal sampling.



c. Scene sampled with exponential foveal geometry.



d. Scene sampled with exponential foveal geometry subdivided by a factor of 4.

Figure 3.2-6. Examples of foveal sampling. (cont'd)

Vision seems to normalize the amount of data necessary to measure a feature attribute. The measurements of large attributes, such as the aspect ratio of a large area, typically require less acuity than small features (Figure 1.1-1). Model based vision, for example, decomposes complicated objects into smaller components; the more complex the object, the more components into which it is decomposed [Burns87]. If these features were of uniform scale, then one would only have to use a uniform acuity sensor with some canonical resolution. However, feature scaling is extremely variable. Not only can objects be of different sizes, but they can be at different distances from the vision platform. Thus, the ideal resolution with which to register a feature (neither oversampling nor undersampling it) falls somewhere within a continuum of resolutions. Additional problems with multiple cameras include the registration of frames with different resolution and camera control. The wide variance of rexel sizes in the linear and exponential patterns represents an important advantage over multiresolution sensor implementations with fewer discrete resolutions, such as two collocated uniform focal plane arrays, each with different scaling or different magnification at the optics. For the same reason that one would typically not consider limiting a hierarchical system to two resolution levels, a foveal sensor should not be limited to two acuities.

The undivided exponential frame has an acuity roll-off which may be too severe for many applications. Subdividing the geometry has the effect of making it more uniform, at the expense of a larger sensor frame. Note how the fovea of the linear and subdivided exponential geometries are large enough to properly register the letter on the card. It is very important when designing a foveal system to properly select the acuity roll-off and fovea size. These should match the distribution of relevance in the scene and the size of the more relevant features expected to be resolved, respectively.

3.2.1 Geometric Analysis of Linear Foveal Pattern

The linear pattern is characterized by concentric square rings of square rexels. There is a central closed ring of rexels measuring 1×1 pixels, surrounded by a ring of rexels measuring 2×2 pixels, surrounded by a ring of 3×3 rexels, and so forth. In general, the n'th ring is of width n and is comprised of $4n$ rexels of size $n \times n$ pixels.

The number of rexels in a foveal pattern with m rings, A_r, is given by

$$A_r = \sum_{i=1}^{m} 4i = 2(m^2 + m) \tag{3-1}$$

The linear dimensions (width and height) of the foveal pattern in terms of pixels is

$$m + (m\text{-}1) + (m\text{-}2) + \ldots + 2 + 1 + 1 + 2 + \ldots (m\text{-}2) + (m\text{-}1) + m$$

$$= 2\sum_{i=1}^{m} i = m^2 + m \tag{3-2}$$

and the area of the pattern in pixels is

$$A_p = (m^2 + m)^2 \tag{3-3}$$

The number of rexels in the linear foveal pattern required to cover an area is thus twice the square root of the number of pixels required to cover the same area:

$$A_r = 2(m^2 + m) = 2\sqrt{(m^2 + m)^2} = 2\sqrt{A_p} \tag{3-4}$$

The linear foveal pattern offers a significant reduction in sensor elements and data when compared to a homogeneous lattice of the same field-of-view and maximum resolution. The savings in sensor elements is given by

$$1 - \frac{A_r}{A_p} = 1 - \frac{2(m^2 + m)}{(m^2 + m)^2} = 1 - \frac{2}{m^2 + m} \tag{3-5}$$

which is a monotonically increasing function with FPA size. Table 3.2.1-1 gives representative values of the savings for different FPA sizes.

For currently used fields-of-view, linear foveal sensor frames offer two to three orders of magnitude less data than the corresponding uniform frames with the same field-of-view and maximum resolution. Figure 3.2.1-1 presents a uniform frame with the same number of samples as Figure 3.2-6b. If the machine task were to identify the type of card (register the letter), data from the uniform sensor would be of little use. This illustrates the effect of improperly matched sensor acuity and relevant feature scale. Under normalization of data size and field-of-view, equation (3-4) relates the sample density at the fovea A_p to the uniform frame sample density A_r. Specifically, extending the number of rings of a

linear pattern such that it has N^2 pixels and then scaling the pattern to an $N \times N$ pixel field-of-view produces a resolution at the fovea a factor $\frac{N}{2}$ greater than that of a pixel.

If a uniform acuity sensor were constructed with the same resolution as in the fovea of a linear foveal pattern, keeping the number of sensor elements constant, the field-of-view would suffer (Figure 3.2.1-2). Under this normalization, (3-4) relates the two fields-of-view, where $\sqrt{A_r}$ is now the linear dimension of the uniform system field-of-view, and $\sqrt{A_p}$ is that of the foveal system. Here, the number of rings of a linear pattern are again extended such that it has N^2 pixels, giving it a field-of-view of size $\frac{N^2}{2} \times \frac{N^2}{2}$ pixels as opposed to $N \times N$ for a uniform resolution frame of same maximum acuity and data size. The constrained field-of-view of the uniform system requires much greater scanning of the scene to accomplish the task. The registration of moving objects is severely handicapped, and applications such as uncooperative target tracking may suffer unacceptably frequent loss of track. In effect, one is faced in the uniform sampling case with a fixed spatial resolution-temporal resolution trade-off. Foveal systems relax the trade-off with the additional degree of freedom provided by the selectable acuity roll-off.

The foveal FPA provides the resolution and field-of-view of uniform lattice FPAs with orders of magnitude less sensor elements. Whereas a 4096 × 4096 FPA greatly exceeds the current manufacturing capabilities for IR sensors (256 × 256=65536 sensor elements), an FPA with 8192 elements arranged in the foveal geometry is feasible and can be produced today. Since current techniques *exceed* the manufacturing requirements of foveal FPAs in many aspects, greater resolution and field-of-view is afforded by increasing the number of rexels while maintaining overall system feasibility.

Size of conventional FPA (pixels) A_p	Size of foveal FPA (rexels) A_r	Number of rings in foveal FPA	$\frac{A_p}{A_r}$	savings in data and elements
256 × 256 = 65,536	512	16	128	99.219%
512 × 512 = 262,144	1024	23	256	99.609%
1024 × 1024 = 1,048,576	2048	31	512	99.804%
2048 × 2048 = 4,194,304	4096	45	1024	99.902%
4096 × 4096 = 16,777,216	8192	63	2048	99.976%

Table 3.2.1-1 Sensor element and data savings with linear foveal pattern.

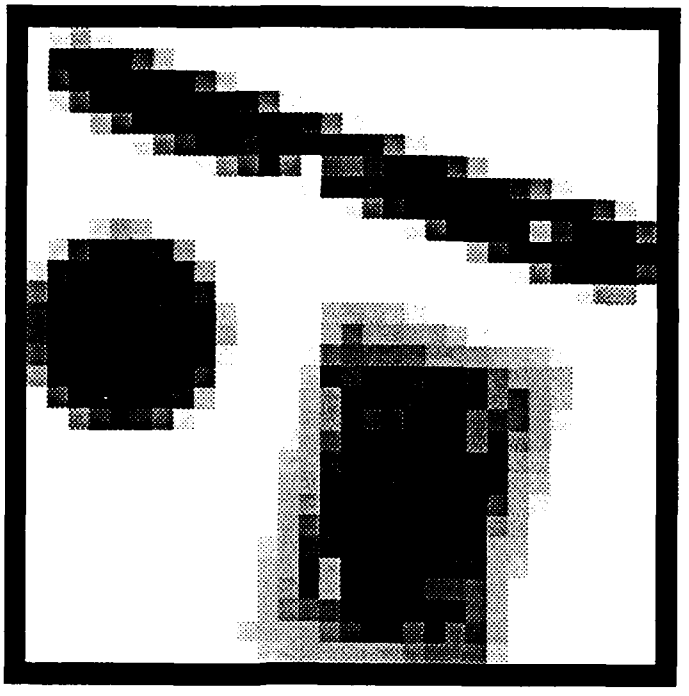


Figure 3.2.1-1. Uniform frame with the data and field-of-view size of Figure 3.2-6b.

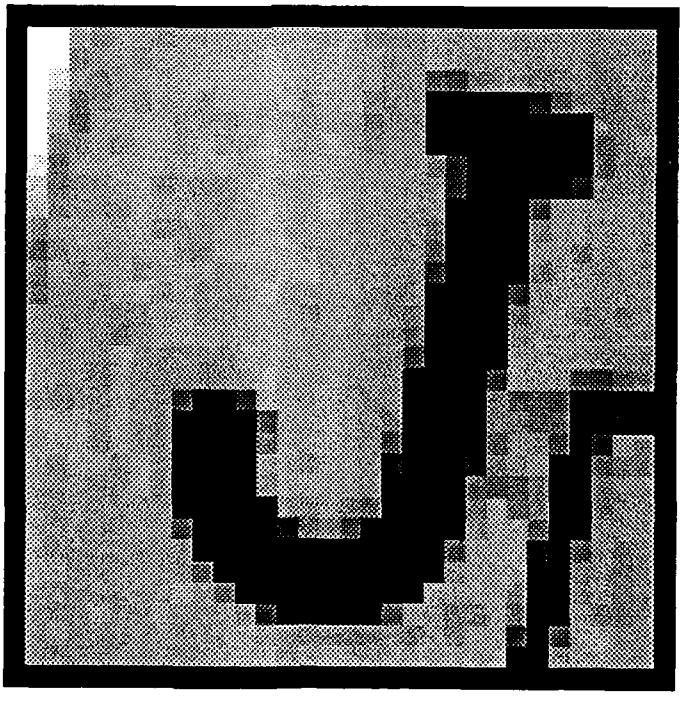


Figure 3.2.1-2. Uniform frame with the data size and maximum resolution of Figure 3.2-6b.

The field-of-view of an FPA is roughly proportional to the size of the FPA (for a given set of front-end optics). Increasing the field-of-view of a conventional FPA by a factor n requires an increase in sensor elements by a factor of n^2:

$$\frac{\text{\# of pixels for extended FOV}}{\text{\# of pixels for unextended FOV}} \propto \frac{N_2 \times N_2}{N_1 \times N_1} = \frac{n \cdot N_1 \times n \cdot N_1}{N_1 \times N_1} = n^2 \qquad (3\text{-}6)$$

Another way to increase the field-of-view is reducing the magnification at the front-end optics or, equivalently, scaling upward the FPA geometry. This approach, however, proportionally reduces resolution and shall not be considered.

The size (field-of-view) of a foveal FPA is increased without sacrificing the resolution at the fovea by adding peripheral rings to the pattern. Let A_r and m be the rexel and ring count of a linear FPA, and A_{rexp} and m_{exp} be the rexel and ring count of a linear foveal FPA with a field-of-view a factor of n greater. The linear dimensions (in pixels) of the foveal FPAs are related by the factor n:

$$m_{exp}^2 + m_{exp} = n\left(m^2 + m\right) \qquad (3\text{-}7)$$

However, the additional field-of-view requires an increase in rexels of only a factor of n as opposed to n^2 pixels for conventional FPAs:

$$\frac{A_{rexp}}{A_r} = \frac{2\left(m_{exp}^2 + m_{exp}\right)}{2\left(m^2 + m\right)} = \frac{2n\left(m^2 + m\right)}{2\left(m^2 + m\right)} = n \qquad (3\text{-}8)$$

Doubling the field-of-view of a linear foveal FPA results in a doubling of the rexel data to be processed, whereas the same improvement in conventional FPA capability results in a quadruple increase in pixel data. Since one increase is linear and the other is exponential, the difference become more significant with greater performance improvement (i.e. a factor 10 improvement generates a factor 10 more rexels or a factor 100 more pixels).

Resolution is improved by proportionally scaling downward the FPA geometry (smaller pixels and rexels). Additional sensor elements must be added to maintain FPA linear dimensions and field-of-view constant. Thus, a factor r improvement in resolution also requires an increase of r rexels or r^2 pixels.

Foveal FPA lithography and read-out circuitry can be standardized for a rexel ring. Thus, one measure of foveal FPA manufacturing complexity is the number of rings required in the implementation. The number of rings m required to cover an area A_p is obtained by solving for m in the expression $A_p = (m^2+m)^2$:

$$m^2 + m - \sqrt{A_p} = 0 \tag{3-9}$$

$$m = \frac{-1 \pm \sqrt{1+4\sqrt{A_p}}}{2} \tag{3-10}$$

Since A_p is positive, m will always be real. Since m must be a positive integer, the "+" term and the nearest integer larger than the value computed by (3-10) is employed, giving

$$m = \left\lceil \frac{-1+\sqrt{1+4\sqrt{A_p}}}{2} \right\rceil \tag{3-11}$$

The double radical is indicative of the small number of rings and rexels required to build a linear foveal FPA. Table 3.2.1-1 gives ring count values for FPAs with various normalized areas. Note that even long term multimillion element requirements are satisfied with foveal geometries containing only a few dozen rings.

3.2.2 Geometric Analysis of Exponential Foveal Pattern

The exponential pattern is characterized by a foveal core of four (two by two) unit sized rexels surrounded by concentric square rings of square rexels. The size of the rexels grows by a multiplicative factor of two, unlike the linear foveal pattern whose rexels grow linearly. The first ring is composed of rexels measuring 1 x 1 pixels, surrounded by a ring of rexels measuring 2 x 2 pixels, surrounded by a ring of 4 x 4 rexels, and so forth. In general, the n'th ring is of width 2^{n-1} and is comprised of 12 rexels of size $2^{n-1} \times 2^{n-1}$ pixels.

The number of rexels in a foveal pattern with m rings, A_r, is given by

$$A_r = 4 + \sum_{i=0}^{m-1} 12 = 4 + 12m \tag{3-12}$$

The linear dimensions (width and height) of the foveal pattern in terms of pixels is

$$2^{m-1} + 2^{m-2} + 2^{m-3} + \ldots + 2 + 1 + 2 \text{ (nucleus)} + 1 + 2 + \ldots 2^{m-3} + 2^{m-2} + 2^{m-1}$$

$$= 2\left(1 + \sum_{i=0}^{m-1} 2^i\right) = 2 + 2\left(\frac{1-2^m}{1-2}\right) = 2^{m+1} \tag{3-13}$$

and the area of the pattern in pixels is

$$A_p = \left(2^{m+1}\right)^2 = 2^{2m+2} = 4 \times 2^{2m} \tag{3-14}$$

The number of rexels required to cover an area is approximately proportional to the logarithm of the number of pixels required to cover the same area:

$$A_r \approx 12 + 12m = 6\log_2 A_p \tag{3-15}$$

The exponential foveal pattern offers an even greater reduction in sensor elements than the linear foveal pattern. This is to be expected, since acuity in the former decreases (rexel size increases) at a faster rate in the peripheral regions of the field-of-view, and peripheral rexels cover more space. The savings in sensor elements is given by

$$1 - \frac{A_r}{A_p} = 1 - \frac{4+12m}{4^{m+1}} = 1 - \frac{1+3m}{4^{m+1}} \tag{3-16}$$

which, as with the linear pattern, is also monotonically increases with FPA size. Table 3.2.2-1 gives representative values of the savings for different FPA sizes.

For currently used fields-of-view, exponential foveal sensor frames offer three to five orders of magnitude less data than the corresponding uniform frames with the same field-of-view and maximum resolution. If the uniform frame has the same number of samples and field-of-view as an exponential frame, (3-15) relates the sample density at the fovea A_p to the uniform frame sample density A_r. Extending the number of rings of a linear pattern such that it has N^2 pixels and then scaling the pattern to an $N \times N$ pixel field-of-view produces a resolution at the fovea a factor $\frac{N^2}{2\log_2 N}$ greater than that of a pixel. If, on the other hand, the uniform sensor has the same maximum acuity, (3-15) relates the foveal field-of-view $\sqrt{A_p}$ to the uniform field-of-view $\sqrt{A_r}$. Again, extending the number

of rings of a linear pattern such that it has N^2 pixels provides a field-of-view of size $2^{\frac{N^2}{12}} \times 2^{\frac{N^2}{12}}$ as opposed to $N \times N$.

Let A_p, A_r and m be the area, rexel, and ring count of an exponential foveal FPA, and A_{pexp}, A_{rexp} and m_{exp} be the area, rexel, and ring count of an exponential foveal FPA with a field-of-view a factor of n greater. The linear dimensions of the foveal FPA (in pixels) are related by n, requiring an additional $\log_2 n$ rings:

$$\sqrt{A_{pexp}} = n \times \sqrt{A_p}$$

$$2^{m_{exp}+1} = n \times 2^{m+1}$$

$$m_{exp} + 1 = \log_2 n + m + 1$$

$$m_{exp} = m + \log_2 n \,. \tag{3-17}$$

The increase in rexels is virtually negligible for even large values of n (m typically >> $\log_2 n$):

$$A_{rexp} = 4 + 12 m_{exp} = 4 + 12m + 12\log_2 n \tag{3-18}$$

$$\frac{A_{rexp}}{A_r} = \frac{4 + 12m + 12\log_2 n}{4 + 12m} = 1 + \frac{3\log_2 n}{1+3m} \tag{3-19}$$

The area of the exponential pattern is increased fourfold, without compromising foveal resolution, by simply adding another ring of 12 rexels at the periphery. The number of rings m required to cover an area A_p is obtained by solving for m in (3-14):

Size of conventional FPA (pixels) A_p	Size of foveal FPA (rexels) A_r	Number of rings in foveal FPA	$\frac{A_p}{A_r}$	savings in data and elements
256 × 256 = 65,536	88	7	744.7	99.866%
512 × 512 = 262,144	100	8	2621.4	99.962%
1024 × 1024 = 1,048,576	112	9	9362.3	99.989%
2048 × 2048 = 4,194,304	124	10	33825	99.997%
4096 × 4096 = 16,777,216	136	11	123362	99.999%

Table 3.2.2-1 Sensor element and data savings with exponential foveal pattern.

$$m = \frac{\log_2 A_p}{2} - 1 \qquad (3\text{-}20)$$

The fact that all rexel sizes are related by a power of two may facilitate the manufacturing of the FPA. Specifically, all rings are identical except for a scaling factor which is a power of two. Only a single lithography template for one ring is required to build an exponential foveal pattern of any size. The template is simply applied repeatedly and scaled by 2, which is the most straightforward scaling factor in computerized lithography. The read-out circuitry is likewise simplified by scale invariance. The power of two scaling also facilitates the processing of sensor data using variants of existing hierarchical image processing techniques, as will be shown in Chapter 6.

3.3 Saccadic Foveation in Unresolved Target Localization

The previous section discussed static properties of two machine implementations of foveal sensors. This section presents and compares the effects of different foveal patterns on system dynamics. These dynamics include the expected and worst case amount of data that must be processed to complete the vision task, and the expected number of redirections of the sensor optical axis. It will be shown that while multiple registrations are required to compensate for the decreased acuity in the periphery of the field-of-view, the number of registrations is typically small and the static data reduction advantages are preserved.

The process of foveation is driven by the task being performed and the scene itself. The task employed in this section is the localization of an unresolved target to the maximum system resolution (one pixel), where the scene consists of a single bright pixel against a dark background. It is assumed that rexel value quantization supports a difference between target present and target absent. This task is both analytically tractable and important to many applications [Drumm89]. The foveation strategy attempts to register the target with the fovea in as few foveations as possible. The saccadic gaze control algorithm attempts to register the target with the fovea as follows:

> *Orient the optical axis to the center of the rexel of the last frame registering the brightest value. The task is complete when this rexel is of size 1×1.*

In general, a search process consists of two serially performed operations: interrogating within the field-of-view for targets, and searching outside the field-of-view (but within the field-of-regard) for additional targets. The second operation can be considered as a reiteration of the first operation at some new gaze angle determined by a spotlight search strategy [Duvoi84]. So that saccadic movement is exclusively considered, only the central operation of interrogating within the field-of-view will be addressed, and it will thus be assumed that the target is initially within the FPA's field-of-view.

When comparing the different foveal geometries with the conventional uniform lattice, it will be assumed that all FPAs are normalized to the same field-of-view and maximum resolution (that of one pixel). This permits performance differences to be attributable exclusively to sensor element geometry.

3.3.1 Average Foveation Performance with Linear Pattern

The task of target localization is completed at the first gaze (registration) if the target appears in any of the four rexels in the fovea of the linear pattern. In this case, the location of the target is known to within the maximum resolution of the system, which is an area measuring one by one pixel. Assuming that the target can appear anywhere in the field-of-view of the FPA with uniform probability, it is unlikely that it will be initially detected by the fovea of a large FPA with many rings (recall that the field-of-view of a rexel is proportional to its size). Instead, it is more likely that the target will be detected by one of the larger rexels; the larger a rexel, the greater its probability of detection.

The probability of a target being registered by a rexel of size $n \times n$ pixels in a linear foveal FPA with m rings is given by

$$P_t(n|m) = \frac{\text{area of all rexels of size } n \times n}{\text{total FPA area}} = \frac{4n \times n^2}{(m^2+m)^2} \qquad (3\text{-}21)$$

and the expected linear dimensions of the rexel registering the target is given by

$$E\{n|m\} = \sum_{n=0}^{m} n \times P_t(n|m) = \sum_{n=0}^{m} \frac{4n^4}{\left(m^2+m\right)^2}$$
$$= \frac{4(2m+1)\left(3m^2+3m-1\right)}{30\left(m^2+m\right)} \cong \frac{4}{5}m \qquad (3\text{-}22)$$

where $E\{\bullet\}$ is the expected value operator. Since the dimension of the linear pattern rexel is the same as the ring index, (3-22) identifies the ring (the extent into the periphery) with which the target is expected to initially be detected.

Peripheral detection localizes the target to within the acuity of the larger rexel. This acuity is far less than the maximum system acuity at the fovea. The foveal machine vision system must steer the optical axis of the FPA so that the target is perceived as close to the fovea as possible. For this analysis, the algorithm for gaze control (foveation) will be to redirect the optical axis to the center of the rexel which is registering the highest (brightest) value.

The process of foveation can be considered as one of reducing the ambiguity of target location. Ambiguity is measured here as the linear dimensions of the rexel known to encompass the target. Thus, the initial target location ambiguity before the first sensor registration is (m^2+m), the linear dimension of the field-of-view.[4] Let ε_i be the ambiguity after the i'th registration, and $\bar{\varepsilon}_i$ be the expected ambiguity after the j'th registration. Then,

$$\bar{\varepsilon}_0 = \varepsilon_0 = \left(m^2+m\right) \qquad (3\text{-}23a)$$

$$\bar{\varepsilon}_1 = E\{\varepsilon_1|m\} \qquad (3\text{-}23b)$$

Approximating (3-11) as

$$m \cong \sqrt{A_p} \qquad (3\text{-}24)$$

and expressing (3-21) and (3-22) in terms of the FPA dimensions $s = \sqrt{A_p}$ instead of ring count m gives

$$P_t(\varepsilon_1|s) = \frac{\text{area of all rexels of size } \varepsilon_1 \times \varepsilon_1}{\text{total FPA area}} = \frac{4\varepsilon_1 \times \varepsilon_1^2}{s^2} \qquad (3\text{-}25)$$

[4] Since the target is assumed to be within the field-of-view of the first sensor registration, and other targets outside this field-of-view are not addressed in the foveation performance analysis, the field-of-view and the field-of-regard can be considered the same.

and

$$E\{\varepsilon_1|s\} \cong \sum_{\varepsilon_1=0}^{s/4} \varepsilon_1 P_i(\varepsilon_1|s) = \sum_{\varepsilon_1=0}^{s/4} \frac{4\varepsilon_1^4}{s^2} \cong \frac{4}{5}\sqrt{s} \tag{3-26}$$

However, since $s=\bar{\varepsilon}_0$, we can write

$$\bar{\varepsilon}_1 = E\{\varepsilon_1|\bar{\varepsilon}_0\} \tag{3-27}$$

The first iteration of target localization reduces the expected ambiguity from $\bar{\varepsilon}_0$ to $\bar{\varepsilon}_1$. The second registration need only address the region $\bar{\varepsilon}_1 \times \bar{\varepsilon}_1$, and not the entire FPA field-of-view. The foveation strategy repositions the optical axis to the center of this region. Thus, only the first (most central) m_2 rings covering the region need to be addressed in the analysis of the next foveation, where m_2 is given by

$$m_2 = \left\lceil \frac{-1+\sqrt{1+4\bar{\varepsilon}_1}}{2} \right\rceil \tag{3-28}$$

The *second iteration* of target localization registration can be considered as another first iteration using an FPA with m_2 rings and linear dimensions of $\bar{\varepsilon}_1$. Consequently,

$$\bar{\varepsilon}_2 = E\{\varepsilon_2|\bar{\varepsilon}_1\} \tag{3-29}$$

which gives rise to the following recursive relation for target localization ambiguity:

$$\bar{\varepsilon}_0 = s \tag{3-30}$$

$$\bar{\varepsilon}_i = E\{\varepsilon_i|\bar{\varepsilon}_{i-1}\} \tag{3-31}$$

Table 3.3.1-1 gives the evolution of target localization ambiguity for FPAs with different fields-of-view. Non integer values for ambiguities and ring counts result from the averaging by the expected value operator. The entry "resolved" indicates that no further foveations are required regardless of target location within the field-of-view. It is seen that even for very large fields-of-view, the number of registrations of average ambiguity reduction required to localize a target is between three and four.

A computer simulation was performed to obtain the exact number of linear geometry foveations required to localize a target in a noise-free environment. Figure

3.3.1-1 illustrates this number as a function of target position (±64 pixels along either axis) relative to the optical axis of the first registration. If the target is detected by the fovea in the first registration, no further registrations are required. The foveation strategy directs the optical axis to the center of the rexel detecting the target. Therefore, if the target is located in the center of a non-fovea rexel in the first frame, it will be localized by the second registration. The combination of the recursive foveal strategy and the linear foveal pattern leads to an interesting foveation strategy map. On average, 2.08 registrations were required to localize the target. Figure 3.3.1-2 illustrates the number of foveations required to localize targets in the first quadrant of a 512 × 512 pixel field-of-view. On average, 2.57 registrations were required to localize the target in a 512 × 512 field-of-view.

# of iterations (i)	$\bar{\varepsilon}_i, m_i$ 256 × 256	$\bar{\varepsilon}_i, m_i$ 512 × 512	$\bar{\varepsilon}_i, m_i$ 1024 × 1024	$\bar{\varepsilon}_i, m_i$ 2048 × 2048	$\bar{\varepsilon}_i, m_i$ 4096 × 4096
0	256, 16	512, 23	1024, 32	2048, 45	4096, 64
1	29.94, 4.68	50.92, 6.65	82.84, 8.62	137.2, 11.23	231.6, 14.73
2	5.21, 1.94	8.51, 2.46	12.26, 3.04	17.92, 3.76	26.54, 4.68
3	resolved	2.20, 1.07	2.91, 1.28	3.87, 1.53	5.21, 1.84
4		resolved	resolved	resolved	resolved

Table 3.3.1-1 Expected target localization ambiguity reduction with linear foveal pattern.

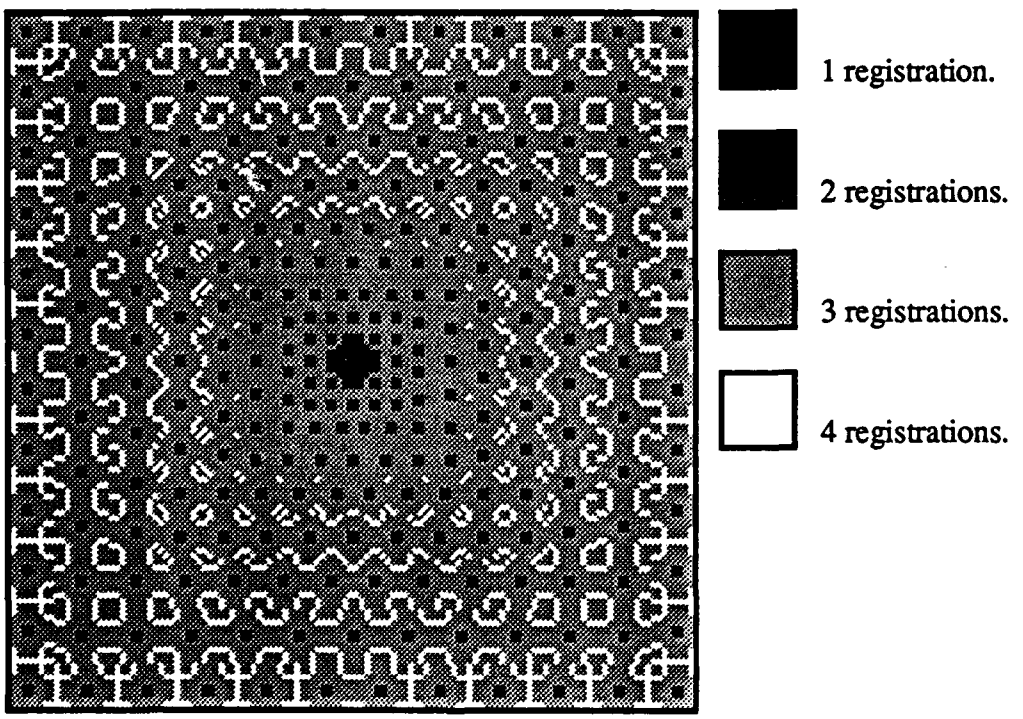


Figure 3.3 1-1. Number of linear foveal pattern registrations required to localize a target as a function of target location (128×128 pixels).

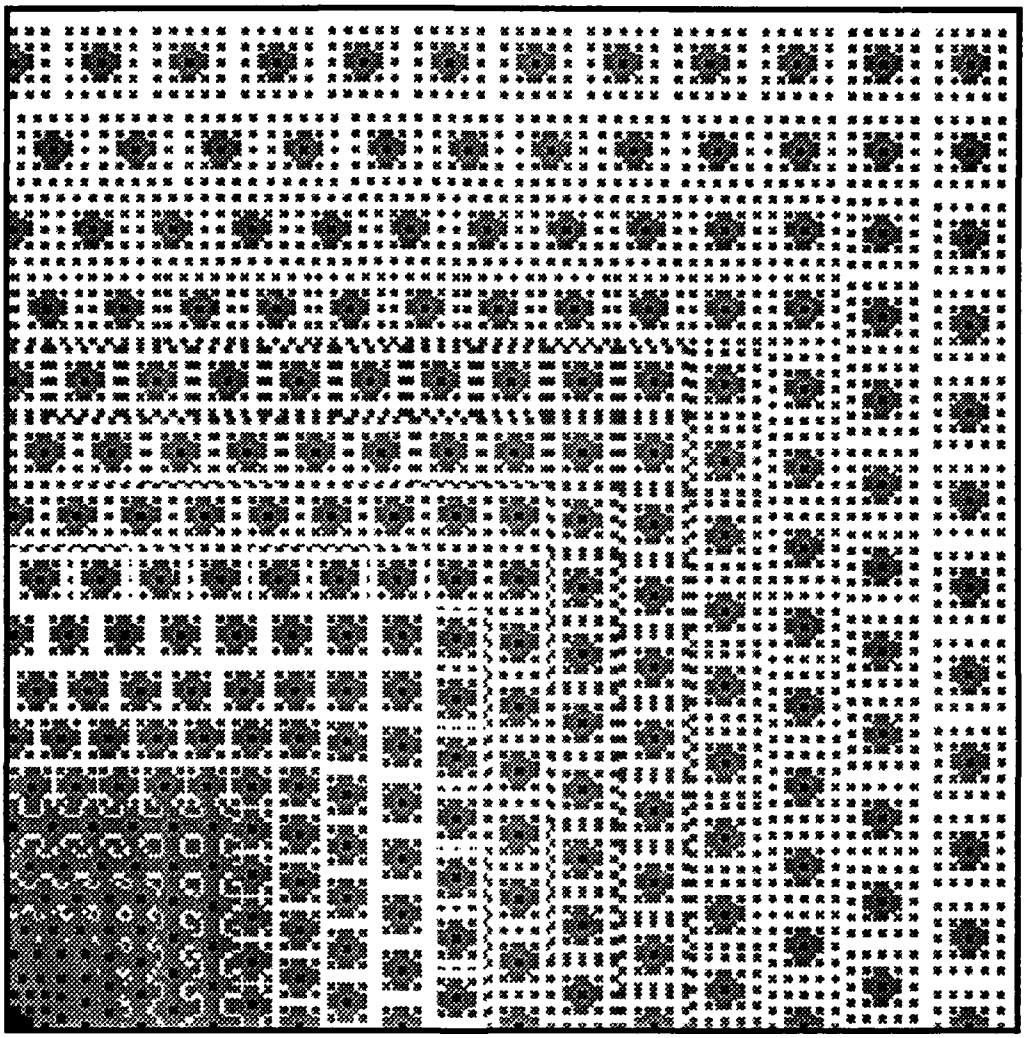


Figure 3.3.1-2. Number of linear foveal pattern registrations required to localize a target as a function of target location (first quadrant of 512×512 pixels). Target positions are with respect to the first quadrant of the initial registration.

An iterative way to compute the average number of foveations n_{av} required to localize a target with the linear pattern is obtained by the following approach. Read the expression

$$\sum_{i=1}^{m} 4i(\ldots) \tag{3-32}$$

as "for every rexel i in a linear foveal frame of size m rings...", where i is the linear dimension of a particular rexel. Similarly, the expression

$$\sum_{i_1=1}^{m} 4i_1\left(\sum_{i_2=1}^{m(i_1)}(\ldots)\right) \tag{3-33}$$

reads "for every rexel i_2 in a linear foveal frame of size $m(i_1)$ rings, resulting from the reduction in ambiguity after one foveation to a rexel of size i_1,...", where i_1 is the linear dimension of a particular rexel in the first frame, i_2 is the linear dimension of a particular rexel in the second frame, and from (3-11),

$$m(i) = \left\lceil \frac{-1+\sqrt{1+4i}}{2} \right\rceil \tag{3-34}$$

The set of indices $I = \{i_1, i_2,\ldots,i_{\max}\}$ obtained from any state of the loop control

$$\sum_{i_1=1}^{m} 4i_1\left(\sum_{i_2=1}^{m(i_1)} 4i_2\left(\cdots\left(\sum_{i_{\max}=1}^{m(i_{\max})} 4i_{\max}(\ldots)\right)\cdots\right)\right) \tag{3-35}$$

denotes the sequence of rexel sizes encountered when localizing a particular pixel in the field-of-regard (the field-of-view of the original registration).[5] The complete execution of the loop control addresses every pixel in the field-of-regard. The number of control recursions, or elements in I, is the worst case foveation count for the particular field-of-regard. This is given by

$$n_{\max,\varepsilon_0} = \log_2(\log_2 \varepsilon_0) + 1. \tag{3-36}$$

[5] The process of localization can be viewed as foveally converging to a single pixel using the described foveation strategy. The set of indices I then represents the state transition of the ambiguity with which the pixel is registered across the sequence of foveal frames.

which is formulated in Section 3.4. Equation (3-34) ensures that the range of the elements of I, $\Re(i)$, are non-zero positive integers which monotonically decrease in value until the value 1 is reached, at which point all remaining elements through to i_{max} are equal to 1:

$$\Re(i_1) \geq \Re(i_2) \geq \cdots \geq \Re\left(i_{n_{\max,\varepsilon_0}}\right) = 1 \tag{3-37}$$

The number of foveations implied by I is easily obtained by analyzing its elements. An element greater than one implies an ambiguity greater than the localization stop rule, so a further foveation is performed. A zero valued element implies that the target has been localized, so no further foveation is required. Consequently, the number of foveations required to localize a target at the location implied by I is

$$n_I = \sum_{j=1}^{n_{\max,\varepsilon_0}} \left[1 - \delta^{(1)}\left(i_j - 1\right)\right] \tag{3-38}$$

where $\delta^{(1)}$ is the unit impulse function

$$\delta^{(1)}(x) = \begin{cases} 1 & x = 0 \\ 0 & x \neq 0 \end{cases} \tag{3-39}$$

The total number of registrations n_t required to localize each pixel in the field-of-regard $\varepsilon_0 \times \varepsilon_0$ is thus

$$n_t = \sum_{i_1=1}^{m} 4i_1 \left(\sum_{i_2=1}^{m(i_1)} 4i_2 \left(\cdots \left(\sum_{i_{n_{\max,\varepsilon_0}}=1}^{m\left(i_{n_{\max,\varepsilon_0}}\right)} 4i_{n_{\max,\varepsilon_0}} \left(\sum_{j=1}^{n_{\max,\varepsilon_0}} \delta^{(1)}\left(i_j - 1\right) \right) \right) \cdots \right) \right) \tag{3-40}$$

where m and $n_{\max,\varepsilon_0}$ are defined by (3-11) and (3-36) respectively. The average number of registrations required to localize a target is simply

$$n_{av} = \frac{n_t}{\varepsilon_0^2} \tag{3-41}$$

3.3.2 Worst Case Foveation Performance with Linear Pattern

The worst case number of foveations required to localize a target occurs when the target is positioned such that in each consecutive refoveation it is detected by the largest rexel in the reduced field-of-view.[6] The linear dimensions of the largest rexel of a linear foveal pattern are equal to the number of rings, or

$$m = \frac{-1\pm\sqrt{1+4\sqrt{A_p}}}{2} = \frac{-1\pm\sqrt{1+4\varepsilon_0}}{2} = \varepsilon_1 \tag{3-42}$$

where $A_p = \varepsilon_0^2$ is the lattice area (in pixels). This can be approximated by (3-24)

$$\varepsilon_1 \cong \frac{\sqrt{4\varepsilon_0}}{2} = \sqrt{\varepsilon_0} \tag{3-43}$$

which provides the recursive relationship

$$\varepsilon_{i+1} = \sqrt{\varepsilon_i} \tag{3-44}$$

$$\varepsilon_i = \underbrace{\sqrt{\sqrt{\cdots\sqrt{\varepsilon_0}}}}_{i \text{ times}} = (\varepsilon_0)^{2^{-i}} \tag{3-45}$$

Let n_{max,ε_0} be the worst case number of registrations required to resolve a target to within one pixel using a linear foveal pattern with linear dimensions of ε_0. Since ε_i equals the linear dimension of a rexel (decreasing in size with increasing i), the ambiguity after the second last foveation will be no greater than 2. Thus,

$$\varepsilon_{n_{max,\varepsilon_0}-1} = (\varepsilon_0)^{2^{-n_{max,\varepsilon_0}-1}} \le 2 \tag{3-46}$$

and in the worst case

$$2^{-n_{max,\varepsilon_0}-1} = \log_{\varepsilon_0} 2 \tag{3-47}$$

From the equality

---

[6] Recall that the active field-of-view can be reduced to the size of the rexel in the last frame known to register the target. Of course, in the presence of strong noise or highly dynamic scenes (multiple moving targets), such a progressive tunnel vision effect would not be advisable.

$$\log_{\varepsilon_0} 2 = \frac{ln2}{ln\varepsilon_0} = \frac{1}{\log_2 \varepsilon_0} \tag{3-48}$$

(3-47) reduces to

$$2^{n_{max,\varepsilon_0}-1} = \log_2 \varepsilon_0 \tag{3-49}$$

and

$$n_{max,\varepsilon_0} = \log_2(\log_2 \varepsilon_0) + 1 \tag{3-50}$$

The worst case number of foveations required to localize a target is small and increases very slowly with increasing field-of-view. Table 3.3.2-1 gives the value of n_{max,ε_0} for a number of different fields-of-view (ε_0). As expected, if the field-of-view is 2×2 pixels, then the pattern consists only of the fovea and the target is instantly localized without redirecting the optical axis. Larger fields-of-view introduce multipixel sized rexels, and axis redirection is required. However, for large FPA sizes, only a small number (four or five) redirections are required in the worst case. Every foveation generates additional data to be processed. However, since the data count of each registration is two to three orders of magnitude less than that from a conventional FPA with the same field-of-view and maximum resolution, substantial savings are maintained.

field-of-view (pixels)	# of registrations (# of foveations + 1)
up to 2 × 2	1
up to 6 × 6	2
up to 42 × 42	3
up to 1806 × 1806	4
up to 4.3M × 4.3M	5

Table 3.3.2-1 Values of n_{max,ε_0} for different linear foveal pattern fields-of-view.

3.3.3 Target Localization with Exponential Foveal Pattern

The static and dynamic properties of the exponential foveal geometry are easier to analyze than the properties of the linear geometry due to the scale invariance of the former (specifically, each ring is 1/4 the size of the next largest ring). The outermost ring of rexels in the exponential pattern occupies 3/4 of the total pattern area. Since the area of consecutive rings are related by a factor of four, the portion of FPA area covered by the i'th largest ring is

$$\frac{\text{area of } i\text{'th largest ring}}{\text{total FPA area}} = \begin{cases} \frac{3}{4} \times \frac{1}{4^i} & i = 0,\ldots,m_0 - 2 \\ \frac{1}{4^i} & i = m_0 - 1 \end{cases} \tag{3-51}$$

where i=0 refers to the largest ring, and m_0 is the total number of rings in the FPA (the fovea cluster of 4×4 rexels of size 1×1 is considered the first ring, i=m_0-1). Expression (3-51) also represents the probability of target detection by the i'th largest ring. Let the event h=k represent detecting the target with the m_0-k'th ring (k'th order detection, or "hit"). Then

$$P(h) = \begin{cases} \frac{3}{4} \times \frac{1}{4^h} & h = 0,\ldots,m_0 - 2 \\ \frac{1}{4^h} & h = m_0 - 1 \end{cases} \tag{3-52a}$$

$$\sum_{h=0}^{m_0-1} P(h) = 1 \tag{3-52b}$$

The event h=k reduces the target location ambiguity by a factor of 0.25^{k+1} to the dimensions of a rexel from the m_0-k'th ring. The next foveation places the optical axis in the center of this rexel. An interesting property stemming from the scale invariance of the exponential foveal geometry is that the coverage of a pattern of m-2 rings, $A_{r,m-2}$, is exactly that of a rexel from the m'th ring, $r_{r,m}$:

$$r_{r,m} = \left(2^{m-1}\right)^2 = 2^{2m-2} \tag{3-53}$$

$$A_{r,m-2} = 4 \times 2^{2(m-2)} = 4 \times 2^{2m-4} = 2^{2m-2} \tag{3-54}$$

Consequently, the number of rings m_1 employed after the event $h=k_1$ is

$$m_1 = m_0\text{-}k\text{-}2. \tag{3-55}$$

Refoveation provides a new event $h=k_2$ which reduces the target location ambiguity to the dimensions of a rexel from the m_1-k_2'th ring from the fovea. The number of rings m_2 employed after the event $h=k_2$ is

$$m_2 = m_1 - k_2 - 2 = m_0 - k_1 - k_2 - 2 - 2 = m_0 - \sum_{n=1}^{2}(k_n + 2) \tag{3-56}$$

where k_n is the order of the n'th detection event. The target location ambiguity after i registrations ε_i is given by

$$\varepsilon_i = 2^{(m_{i-1}-k_i)-1} = 2^{m_i+1} = 2^{\left(m_0-\sum_{n=1}^{i}(k_n+2)\right)+1} = 2^{m_0-2i+1-\sum_{n=1}^{i}k_n} \tag{3-57}$$

Target localization is achieved when $\varepsilon_i=1$, which implies the condition

$$0 = m_0 - 2i + 1 - \sum_{n=1}^{i} k_n \tag{3-58a}$$

$$\sum_{n=1}^{i} k_n = m_0 - 2i + 1 \tag{3-58b}$$

relating detection order, overall field-of-view, and the number of registrations required to unambiguously localize the target.

The expected detection order $\bar{k}$ of a foveation is obtained from the probability of detecting a target with the k'th largest ring:[7]

$$\bar{k} = E\{k\} = \sum_{k=0}^{\infty} k\mathrm{P}(k) = \frac{3}{4}\sum_{k=0}^{\infty} k\left(\frac{1}{4}\right)^k = \frac{1}{3} \tag{3-59}$$

The expected target location ambiguity $\bar{\varepsilon}_i$ is obtained from the expected detection order:

---

[7] For mathematical tractability, it is assumed that there are an infinite number of rings; to preserve realism, detections of order greater than the actual number of physical rings m are mapped into $k=m$-1. This approach to the solution for $\bar{k}$ permits the use of the infinite arithmetic-geometric series without addressing terminal conditions, namely the fovea, which in the strict sense is the one non-scale invariant component of the exponential foveal geometry.

$$\bar{\varepsilon}_{i+1} = 2^{m_{i+1}+1} = 2^{(m_i-\bar{k}-2)+1} = 2^{m_i+1}2^{-\bar{k}-2} = 0.198\varepsilon_i \qquad (3\text{-}60)$$

The average target location ambiguity reduction is very close to the worst case performance, where detection order is constantly zero and ambiguity is reduced by 2. This is because the last ring of the exponential geometry, which generates the zero order detections, occupies most (three fourths) of the FPA area and three times more than all other rings combined.

The number of registrations with expected location ambiguity reduction required to localize the target, $\bar{n}$, is obtained by recursively applying (3-60) $\bar{n}$ times with field-of-view as the initial ambiguity value and 1 as the final ambiguity value, and solving for $\bar{n}$:

$$\bar{\varepsilon}_{\bar{n}} = 0.198^{\bar{n}}\varepsilon_0 = 1 \qquad (3\text{-}61)$$

$$\bar{n} = -\log_{.198}\varepsilon_0 = -\frac{\log_2\varepsilon_0}{\log_2(.198)} = \frac{\log_2\varepsilon_0}{2.336} \qquad (3\text{-}62)$$

Table 3.3.3-1 gives values of $\bar{n}$ for different fields-of-view. Noninteger registration numbers are to be expected, since $\bar{n}$ represents an average value. For large fields-of-view, the number of exponential geometry foveations is greater than with the linear pattern because the acuity in the former decreases faster with distance from the optical axis. The peripheral rexels have greater coverage, providing greater data savings, but also greater ambiguity which must be resolved through additional foveations.[8]

A computer simulation was performed to obtain the exact number of exponential geometry foveations required to localize a target in a noise-free environment. Figure 3.3.3-1 illustrates this number as a function of target position (±64 pixels along either axis) relative to the optical axis of the first registration. On average, 2.12 registrations were required to localize the target. This is similar to the average registration count of Figure 3.3.2-1 because of the small field-of-regard. Figure 3.3.3-2 illustrates the number of foveations required to localize targets in the first quadrant of a 512 × 512 pixel field-of-view. On average, 2.78 registrations were required to localize the target within a 512 × 512 field-of-view. The sharper decrease of exponential pattern resolution mandated one more foveation than with the linear geometry. The experimental values are less than those

[8] The large fovea of the exponential pattern produces fewer foveations than the linear pattern when the field-of-view is relatively small. For large fields-of-view, the effect of large rexels (low acuity) in the periphery outweighs this advantage, and more foveations are required. Different vision tasks, such as those which must resolve objects greater than one pixel in size, can better exploit a larger fovea.

given by the expected ambiguity reduction model because the model employs statistical moments and not the actual probabilities for the foveation sequences, thus not explicitly accounting for low probability detection with small rexels.

field-of-view (pixels)	# of registrations (# of foveations + 1)
up to 16 × 16	1.712
up to 64 × 64	2.568
up to 128 × 128	2.997
up to 256 × 256	3.425
up to 512 × 512	3.853
up to 1024 × 1024	4.281
up to 4096 × 4096	5.137
up to 16384 × 16384	5.993
up to 65536 × 65536	6.849
up to 262K × 262K	7.705

Table 3.3.3-1 Values of $\overline{n}$ for different exponential foveal pattern fields-of-view computed by average ambiguity reduction.

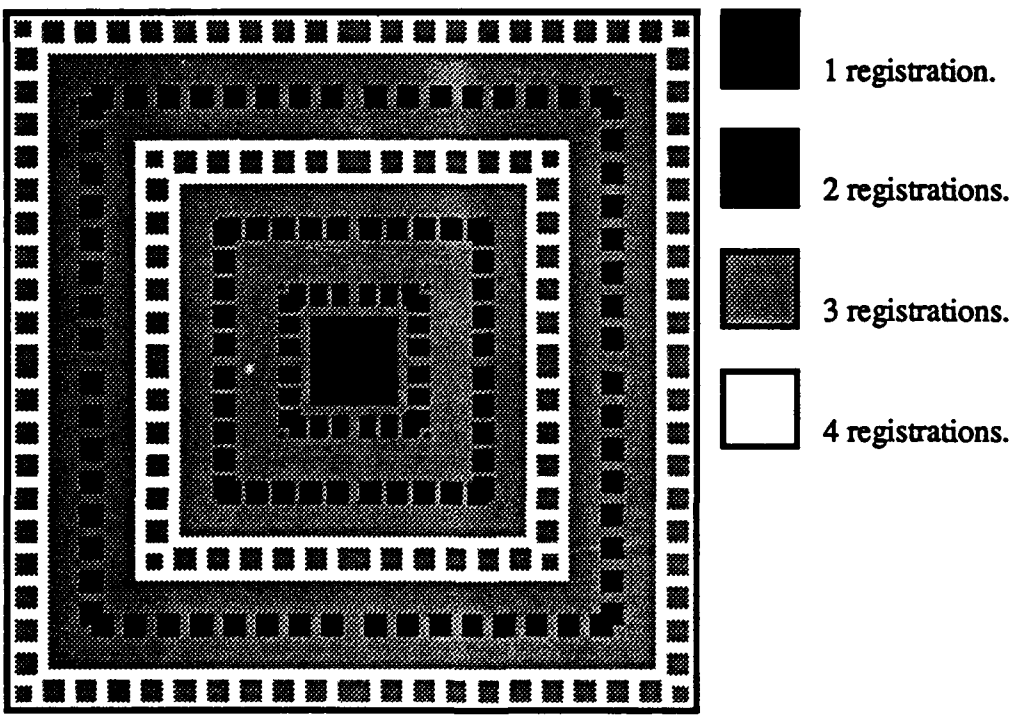


Figure 3.3.3-1. Number of exponential foveal pattern registrations required to localize a target as a function of target location (128×128 pixels).

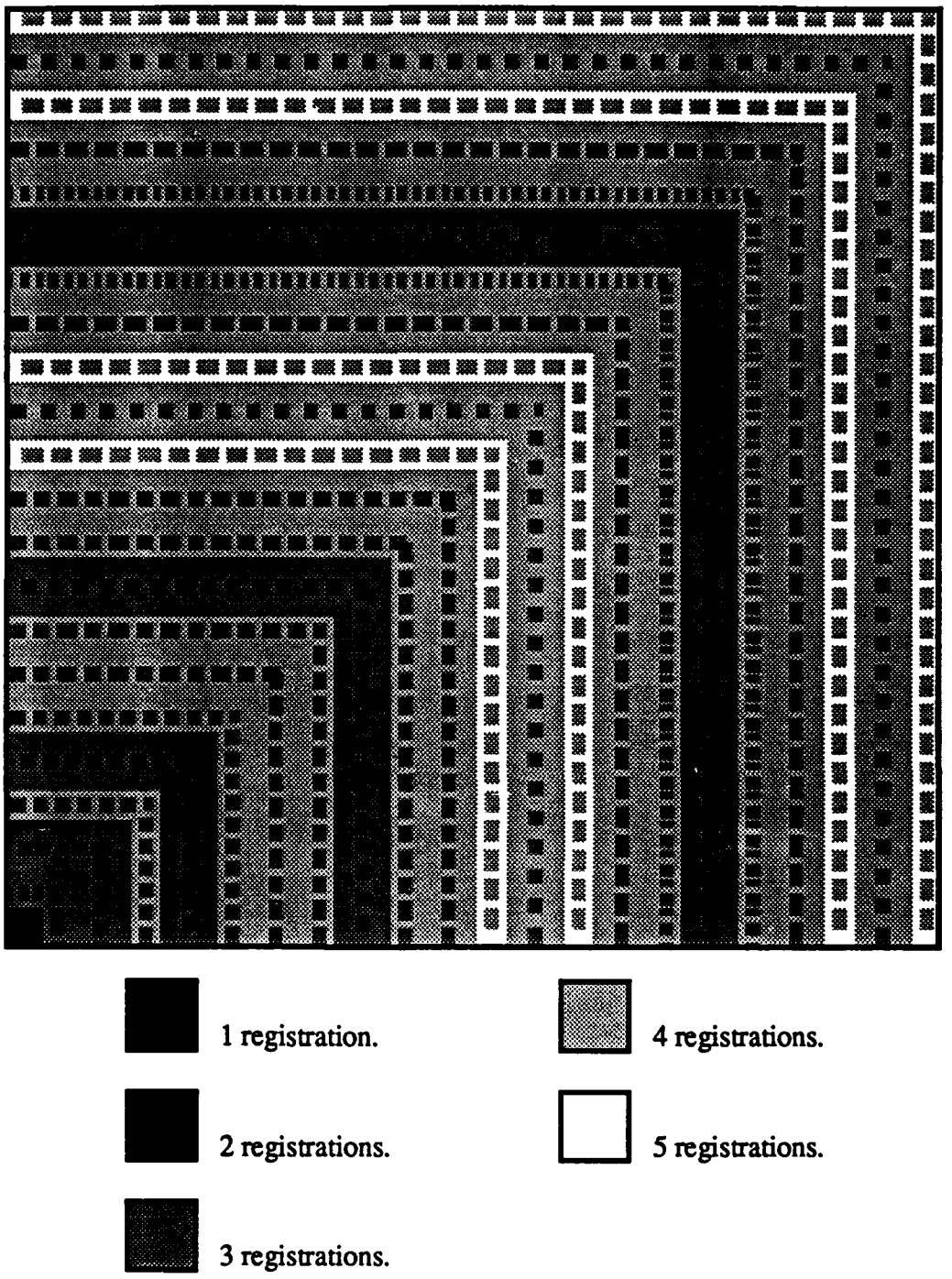


Figure 3.3.3-2. Number of exponential foveal pattern registrations required to localize a target as a function of target location (first quadrant of 512×512 pixels). Target positions are with respect to the first quadrant of the initial registration.

Another way to estimate the average number of foveations required to localize an unresolved target against a dark background with the exponential pattern is to employ the probabilities of hit sequences as opposed to the expected ambiguity reduction. Consider a sequence of n hits $H=\{h_1,h_2,...,h_n\}$, where h_i is the order of the i'th hit. It is seen from (3-58) that the order of elements in a hit sequence does not affect the result of the overall series of foveations. For example, three foveations of order 0 followed by a hit of order 1, $H=\{0,0,0,1\}$, results in the same reduction in ambiguity as one hit of order 0 followed by one hit of order 1 and two more of order 0, $H=\{0,1,0,0\}$. What distinguishes the entropy reduction of one sequence from another is the total number of hits.

Let x_k be the number of hits of order k in a given foveation sequence. The target localization performance of the sequence is completely characterized by the set of hit order counts $X = \{x_0,x_1,...,x_{m_0-1}\}$. Note that there cannot be a hit of order greater than m_0-1 because there are no rings in the pattern to support such a hit, and

$$\sum_{i=0}^{m_0-1} x_i = n \tag{3-63}$$

where n is the overall number of hits (registrations). The condition for target localization (3-58) can be rewritten in terms of X from the relation

$$\sum_{i=0}^{m_0-1} ix_i = \sum_{j=0}^{n} h_j \tag{3-64}$$

giving

$$\sum_{i=0}^{m_0-1} ix_i = m_0 - 2n + 1 \tag{3-65}$$

The probability of obtaining a foveation sequence characterized by X is given by the multinomial distribution

$$P(X) = P\left(x_0,x_1,\ldots,x_{m_0-1}\right) = \frac{n}{x_0!x_1!\cdots x_{m_0-1}!}\theta_0^{x_0}\theta_1^{x_1}\cdots\theta_{m_0-1}^{x_{m_0-1}} \tag{3-66}$$

where $\theta_h=P(h)$ is the probability of obtaining a hit of order h as given by (3-52). Combining these expressions gives

$$P(X) = P\left(x_0, x_1, \ldots, x_{m_0-1}\right) = \frac{n}{x_0!x_1!x_2!\cdots x_{m_0-2}!x_{m_0-1}!} \times$$

$$\left(\frac{3}{4}\right)^{x_0} \times \left(\frac{3}{4}\right)^{x_1}\left(\frac{1}{4}\right)^{x_1} \times \left(\frac{3}{4}\right)^{x_2}\left(\frac{1}{4^2}\right)^{x_2} \times \cdots \times \left(\frac{3}{4}\right)^{x_{m_0-2}}\left(\frac{1}{4^{m_0-2}}\right)^{x_{m_0-2}} \times \left(\frac{1}{4^{m_0-1}}\right)^{x_{m_0-1}} \quad (3\text{-}67)$$

which upon factorizing reduces to

$$P(X) = P\left(x_0, x_1, \ldots, x_{m_0-1}\right) = \frac{n}{x_0!x_1!x_2!\cdots x_{m_0-2}!x_{m_0-1}!}\left(\frac{3}{4}\right)^{n-x_{m_0-1}}\left(\frac{1}{4}\right)^{\alpha} \quad (3\text{-}68)$$

where

$$\alpha = \sum_{i=0}^{m_0-1} i x_i = \sum_{j=0}^{n} h_j \quad (3\text{-}69)$$

The expected number of registrations can be obtained by solving for the expected hit order set

$$\overline{X} = E\{X\} = \left\{\bar{x}_0, \bar{x}_1, \ldots, \bar{x}_{m_0-1}\right\} \quad (3\text{-}70)$$

and using (3-65) to obtain the expected number of registrations:

$$\bar{n} = \frac{m_0 + 1 - \sum_{i=0}^{m_0-1} i\bar{x}_i}{2} \quad (3\text{-}71)$$

The expected number of hits of order i from a sequence of n hits can be approximated by

$$\bar{x}_i \cong n\theta_i^{x_i} = \begin{cases} n \times \frac{3}{4} \times \frac{1}{4^i} & i = 0, \ldots, m_0 - 2 \\ n \times \frac{1}{4^i} & i = m_0 - 1 \end{cases} \quad (3\text{-}72)$$

The condition for target localization is then expressed in terms of m_0, n and θ_h by

$$\sum_{i=0}^{m_0-1} i\bar{x}_i = (m_0-1)\bar{x}_{m_0-1} + \sum_{i=0}^{m_0-2} i\bar{x}_i = (m_0-1)n\theta_{m_0-1} + \sum_{i=0}^{m_0-2} in\theta_i$$

$$= (m_0-1)n\left(\frac{1}{4}\right)^{m_0-1} + \frac{3n}{4}\sum_{i=0}^{m_0-2} i\left(\frac{1}{4}\right)^i$$

$$= (m_0-1)n\left(\frac{1}{4}\right)^{m_0-1} + \frac{3n}{4}\frac{\frac{1}{4}\left[1-(m_0-1)\left(\frac{1}{4}\right)^{m_0-2}+(m_0-2)\left(\frac{1}{4}\right)^{m_0-1}\right]}{\left(1-\frac{1}{4}\right)^2}$$

$$= \frac{n}{3}\left[1-\left(\frac{1}{4}\right)^{m_0-1}\right] \tag{3-73}$$

Equating this with m_0-2n+1 and solving for n gives the expected number of registrations as a function of ring count

$$\bar{n} = \frac{m_0+1}{2+\frac{1}{3}\left[1-\left(\frac{1}{4}\right)^{m_0-1}\right]} \cong \frac{3}{7}(m_0+1) \tag{3-74}$$

Equivalently, using the relationship between ring count, exponential foveal geometry FPA area (3-20), and initial ambiguity (3-30),

$$m_0 = \frac{\log_2 A_p}{2} - 1 = \log_2 \varepsilon_0 - 1 \tag{3-75}$$

the expected number of registrations is expressed in terms of ambiguity:

$$\bar{n} = \frac{\log_2 \varepsilon_0}{2+\frac{1}{3}\left[1-\frac{16}{\varepsilon_0^2}\right]} \cong \frac{3}{7}\log_2 \varepsilon_0 \tag{3-76}$$

This expression produces lower values than equation (3-62) because it better accounts for the low probability but nevertheless possible high order hits, such as when the target is at the fovea of the initial registration. Table 3.3.3-2 gives the expected number of registrations for different fields-of-view using the above expected hit order set model. Note that the values are closer to the measured data than those generated by the average ambiguity reduction model.

field-of-view (pixels)	# of registrations (# of foveations + 1)
up to 16 × 16	1.730
up to 64 × 64	2.573
up to 128 × 128	3.000
up to 256 × 256	3.429
up to 512 × 512	3.857
up to 1024 × 1024	4.286
up to 4096 × 4096	5.143
up to 16384 × 16384	6.000
up to 65536 × 65536	6.857
up to 262K × 262K	7.729

Table 3.3.3-2 Values of $\bar{n}$ for different exponential foveal pattern fields-of-view computed by expected hit order set.

3.3.4 Worst Case Foveation Performance with Exponential Pattern

As with the linear geometry, the worst case number of foveations required to localize a target occurs when the target is positioned such that in each consecutive refoveation it is detected by the largest rexel within the reduced field-of-view. In this case, all detections are of order k=0, and

$$\varepsilon_{i+1} = \frac{\varepsilon_i}{4} \tag{3-77}$$

$$m_{i+1} = m_i\text{-}2 \tag{3-78}$$

$$\varepsilon_n = \frac{\varepsilon_0}{4^n} \tag{3-79}$$

$$m_n = m_0\text{-}2n. \tag{3-80}$$

The worst case number of foveations n_{max} to localize a target ($m_{n_{max}}$=1) is approximately equal to half the number of rings in the pattern:

$$n = \frac{m_0\text{-}m_n}{2} \tag{3-81}$$

$$n_{max} = \frac{m_0 - 1}{2} = \frac{\log_2 \varepsilon_0}{2} \tag{3-82}$$

The worst case number of foveations required to localize a target is small and increases very slowly with field-of-view. Table 3.3.4-1 gives the value of n_{max,ε_0} for a number of different fields-of-view (ε_0). As expected, for large fields-of-view, the worst case number of exponential lattice foveations is greater than with the linear lattice. Also, since the expected detection order is close to the worst case of zero, the number of foveations of expected ambiguity reduction necessary to localize the target is only marginally better than the worst case performance. This is not to say that the performance of the exponential foveal system is inferior to the linear; it indicates that system performance is more consistent.

field-of-view (pixels)	# of registrations (# of foveations + 1)
up to 16 × 16	2
up to 64 × 64	3
up to 256 × 256	4
up to 1024 × 1024	5
up to 4096 × 4096	6
up to 16384 × 16384	7
up to 65536 × 65536	8
up to 262K × 262K	9

Table 3.3.4-1. Values of n_{max,ε_0} for different exponential pattern fields-of-view.

3.3.5 Summary of Linear and Exponential Foveation Performance

It is shown that with either geometry, the number of foveations required to localize an unresolved target is very small. The linear pattern requires fewer frames than the exponential pattern, specifically three or four for conventional sensor fields-of-view as opposed to four or five. The exponential pattern requires more frames because its more pronounced acuity roll-off results in a greater initial average target position ambiguity and and because convergence is slower; a foveation reduces ambiguity by a minimum factor of

four with the exponential pattern, whereas with the linear pattern, ambiguity is reduced by the square root. However, its larger fovea permits the exponential pattern to localize the target at greater foveation miss distances.

The similarity in foveation count for both patterns implies that for this task there does not exist a proportional trade-off between frame size (in rexels) and foveation count; the product of frame size and the number of frames processed is much smaller in the exponential case. This behavior of foveal spatiotemporal resolution allocation can be exploited by using smaller bandwidth hardware in the implementation of foveal systems employing the exponential pattern. The next section compares the bandwidth of various different systems performing target localization.

An interesting fact is that the sequence of saccades with the linear pattern resembles the sequence of saccades by the human eye performing the same task. The biological system performs an initial saccade which undershoots a stationary target by 1% to 5% and is then followed by a small number of "corrective" saccades [Lemij89]. The worst case target location ambiguity after the first foveation, given by (3-43), can be used as an upper limit on the targeting error of the initial machine saccade. For fields-of-view typical of machine vision, the targeting error is on the order of 1% to 2%. The consecutive saccades "home-in" on the target, as illustrated by Figures 3.3.1-1 and 3.3.1-2.

3.4 Comparison of Localization Performance with Conventional Machine Vision Systems

In this section, we compare the performance of unresolved target localization with the linear and exponential pattern foveal machine vision systems with that of three types of current machine vision systems. The three current systems employ conventional uniform resolution FPAs and differ in their processing hardware. All systems are described below:

1. Uniprocessor linear pattern foveal machine vision system: a single processor sequentially processes each rexel from a linear FPA frame in one time step.

2. Multiprocessor linear pattern foveal machine vision system: a non-von Neumann array of processors processes all rexels of a frame from a linear FPA frame in one time step.

3. Uniprocessor exponential pattern foveal machine vision system: a single processor sequentially processes each rexel from an exponential FPA frame in one time step.

4. Multiprocessor exponential pattern foveal machine vision system: a non-von Neumann array of processors processes all rexels from an exponential FPA frame in one time step.

5. Uniprocessor uniresolution machine vision system: a single processor sequentially processes each pixel from a uniform resolution FPA in one time step.

6. Multiprocessor uniresolution machine vision system: an array of processors processes all pixels in a frame from a uniform resolution FPA in one time step.

7. Pyramid machine vision system: a non-von Neumann architecture generates a multiresolution multilevel image pyramid from each frame of an uniresolution FPA, and processes one pyramid level in one time step.

The attributes compared are the frame size, which is analogous to system hardware (e.g., frame buffer size, and processor count in the case of multiprocessor systems), and the time steps required to complete the task. The product of these two attributes is used as a global (area×time) measure of computational complexity. This number is typically used as an inverse figure of merit, where smaller values indicate greater feasibility or desirability.

3.4.1 Uniprocessor Linear Pattern Foveal System

This system processes multiple frames from a linear foveal FPA to localize the target. The amount of data per frame is $2\varepsilon_0$ (rexel) values, where ε_0 is the linear dimension of the field-of-view measured in normalized pixels. One value is processed by the processor in one time step. This system requires approximately 3.5 frames (registrations at different optical axis pointing angles) on average to localize the target in typical and large fields-of-view (512×512 and 4096×4096) (Table 3.3.1-1). The worst case number of frames required for target localization is $\log_2(\log_2\varepsilon_0)+1$.

3.4.2 Multiprocessor Linear Pattern Foveal System

This system also processes multiple frames from a linear foveal FPA to localize the target. All $2\varepsilon_0$ (rexel) values per frame are processed in one time step by a single-instruction-multiple-data (SIMD) multiprocessor array. An additional $2\sqrt{2\varepsilon_0}$ steps per frame (twice the square root of the frame size, for a horizontal neighbor transfer of data to a single column accumulator, followed by a vertical transfer of data up the accumulator column) are required to flush the result out of the system.[9] Array set-up can be performed

[9] This is the typical number of data flush steps required by a 4-neighbor connected (uniform Cartesian lattice) systolic array. The foveal geometry is not uniform, but as long as the rexels are ordered, the SIMD architecture need not be topologically identical to the sensor (e.g., a linear array of processors could be employed). The higher the dimensionality of the SIMD architecture topology, the lower the data flush overhead; specifically, the overhead is $n\sqrt[n]{D}$ where n is the dimensionality of the architecture and D is the total data amount. A 2-dimensional systolic array is employed because it is prevalent in many applications.

concurrently, and thus introduces no additional overhead. This system requires approximately 3.5 frames on average to localize the target. The worst case number of frames required for target localization is $\log_2(\log_2\varepsilon_0)+1$.

3.4.3 Uniprocessor Exponential Pattern Foveal System

This system processes multiple frames from an exponential foveal FPA to localize the target. The amount of data per frame is $6\log_2\varepsilon_0$ (rexel) values. One value is processed by the single processor in one time step. This system requires approximately $0.43\times\log_2(\varepsilon_0)$ - 0.86 frames on average to localize the target. The worst case number of frames required for target localization is $0.50\times\log_2(\varepsilon_0)$.

3.4.4 Multiprocessor Exponential Pattern Foveal System

This system also processes multiple frames from an exponential foveal FPA to localize the target. All $6\log_2\varepsilon_0$ (rexel) values per frame are processed in one time step by a SIMD multiprocessor array. An additional $2\sqrt{6\log_2\varepsilon_0}$ steps per frame are required to flush the result out of the system. Array set-up can be performed concurrently, and thus introduces no additional overhead. This system requires approximately $0.43\times\log_2(\varepsilon_0)$ - 0.86 frames on average to localize the target. The worst case number of frames required for target localization is $0.50\times\log_2(\varepsilon_0)$.

3.4.5 Uniprocessor Uniresolution Machine Vision System

This system processes a single frame of data from an FPA with uniform resolution throughout its field-of-view. The amount of data per frame is ε_0^2 (pixel) values. The system must process the entire frame on average and in the worst case to localize the target (find the brightest pixel). One value is processed by the single processor in one time step.

The uniprocessor uniresolution system is the most prevalent in the field. It is also the slowest of the architectures discussed in this comparison. Because uniresolution image processing algorithms are easily implemented on a multiprocessor system [Dough87], a small number of processors (on the order of ten) are sometimes employed. This has the effect of almost proportionally decreasing the number of time steps required by the system to process a frame of data. However, it is shown that even a factor of ten improvement in speed does not compensate for the inherently enormous amount of data generated by the uniform resolution FPA.

3.4.6 Multiprocessor Uniresolution Machine Vision System

This system also processes a single frame of data from an FPA with uniform resolution. The amount of data per frame is ε_0^2 (pixel) values. The system must process the entire frame on average and in the worst case to localize the target. All values are processed in one time step by a SIMD multiprocessor array. In order to flush results out of the array, $2\varepsilon_0$ additional steps per frame are required. Array set-up can be performed concurrently, and thus introduces no additional overhead. This architecture can be unfeasible for large (greater than 512×512) processor arrays.

3.4.7 Pyramid Machine Vision System

This system processes a single frame of data from an FPA with uniform resolution. It generates a Gaussian pyramid (four sibling nodes to one parent node) with $\frac{4}{3}\varepsilon_0^2$-$\frac{1}{3}$ (pixel) values [Burt84]. The SIMD system requires $2\times\log_2(\varepsilon_0)$ time steps to process the frame [Blanf88]: $\log_2(\varepsilon_0)$ steps to generate in a bottom-up fashion the pyramid data structure from the frame data, and another $\log_2(\varepsilon_0)$ steps to conduct a top-down analysis. The top-down analysis automatically flushes the result. However, pyramid set-up cannot be performed concurrently with top-down analysis. Thus, the $2\times\log_2(\varepsilon_0)$ time steps do not

include frame loading time. This architecture can be unfeasible for pyramids with large (greater than 512×512) base processor arrays.

3.4.8 Average Performance Comparison

The average performance of a machine vision system is defined here by two parameters: the expected number of computational time steps T_{av} required to localize an unresolved target, and the amount of required frame memory A_m. For multiprocessor systems, T_{av} is the product of expected number of registrations with the number of time steps required to retrieve the result of frame processing from the processor array; all the data in a frame is processed in parallel in one time step. For uniprocessor systems, T_{av} is the product of expected number of registrations with the data per registration (data per frame), since each frame datum is processed sequentially, one per time step. The A_m parameter represents the system frame buffer; in the case of the pyramid, A_m is the size of the hierarchical data structure. The average computational complexity C_{av} of a machine vision system is computed here as the product of T_{av} and A_m.

Table 3.4.8-1 presents the analytical expressions for the average performance and computational complexity of the previously described machine vision systems. The systems are normalized to have the same maximum resolution (one pixel) and field-of-view ($\varepsilon_0 \times \varepsilon_0$). This way, none is handicapped and the differences in expressions are attributable exclusively to the approach of the system to the task of target localization. Tables 3.4.8-2 through 3.4.8-4 present numerical values from these expressions for the fields-of-view of 512×512, 1024×1024, and 4098×4096 respectively.

For the multiprocessor systems, the number of processors is equal to A_m. At large frame sizes, the uniform acuity multiprocessor and pyramid systems are unfeasible because of the large number of processors required. The multiprocessor foveal systems, however, require several orders of magnitude less processors (due to the smaller data count).

Machine Vision System Type	T_{av}	A_m	C_{av}
linear foveal pattern uniprocessor	$7\varepsilon_0$	$2\varepsilon_0$	$14\varepsilon_0^2$
linear foveal pattern multiprocessor	$3.5\times2\sqrt{2\varepsilon_0}$	$2\varepsilon_0$	$14\varepsilon_0\times\sqrt{2\varepsilon_0}$
exponential foveal pattern uniprocessor	$2.58[\log_2^2(\varepsilon_0)-2\log_2(\varepsilon_0)]^2$	$6\log_2\varepsilon_0$	$15.58[\log_2^3(\varepsilon_0)-2\log_2^2(\varepsilon_0)]$
exponential foveal pattern multiprocessor	$0.43\times[\log_2(\varepsilon_0)-2]\times 2\sqrt{6\log_2\varepsilon_0}$	$6\log_2\varepsilon_0$	$5.16[\log_2^2(\varepsilon_0)-2\log_2(\varepsilon_0)]\sqrt{6\log_2\varepsilon_0}$
uniform acuity uniprocessor	ε_0^2	ε_0^2	ε_0^4
uniform acuity multiprocessor	$2\varepsilon_0$	ε_0^2	$2\varepsilon_0^3$
multiprocessor pyramid	$2\times\log_2(\varepsilon_0)$+ frame load time	$\frac{4}{3}\varepsilon_0^2$	$\frac{8}{3}\varepsilon_0^2\times\log_2(\varepsilon_0)$

Table 3.4.8-1 Analytical expressions for average performance and computational complexity of machine vision systems performing target localization.

Machine Vision System Type	T_{av}	A_m	C_{av}
linear foveal pattern uniprocessor	3584	1024	3.7×10^6
linear foveal pattern multiprocessor	224	1024	230×10^3
exponential foveal pattern uniprocessor	162	54	8.8×10^3
exponential foveal pattern multiprocessor	44.2	54	2.4×10^3
uniform acuity uniprocessor	131,072	262,144	69×10^9
uniform acuity multiprocessor (unfeasible)	1024	262,144	270×10^6
multiprocessor pyramid (unfeasible)	18+ set-up time	349,525	6.3×10^6

Table 3.4.8-2 Average performance and computational complexity of machine vision systems with a field-of-view of 512×512 performing target localization.

Machine Vision System Type	T_{av}	A_m	C_{av}
linear foveal pattern uniprocessor	7168	2048	15×10^6
linear foveal pattern multiprocessor	316.8	2048	650×10^3
exponential foveal pattern uniprocessor	206	60	12×10^3
exponential foveal pattern multiprocessor	53.3	60	3.2×10^3
uniform acuity uniprocessor	1.05×10^6	1.05×10^6	1.1×10^{12}
uniform acuity multiprocessor (unfeasible)	2048	1.05×10^6	2.2×10^9
multiprocessor pyramid (unfeasible)	20+ set-up time	1.40×10^6	28×10^6

Table 3.4.8-3 Average performance and computational complexity of machine vision systems with a field-of-view of 1024x1024 performing target localization.

Machine Vision System Type	T_{av}	A_m	C_{av}
linear foveal pattern uniprocessor	28.7×10^3	8192	230×10^6
linear foveal pattern multiprocessor	633.6	8192	5.2×10^6
exponential foveal pattern uniprocessor	309.6	72	22×10^3
exponential foveal pattern multiprocessor	73.0	72	5.3×10^3
uniform acuity uniprocessor	16.8×10^6	16.8×10^6	280×10^{12}
uniform acuity multiprocessor (unfeasible)	8192	16.8×10^6	140×10^9
multiprocessor pyramid (unfeasible)	24+ set-up time	22.4×10^6	540×10^6

Table 3.4.8-4 Average performance and computational complexity of machine vision systems with a field-of-view of 4096x4096 performing target localization.

3.4.9 Worst Case Performance Comparison

The worst case performance of a machine vision system is defined by two parameters: the worst case number of time steps T_{max} required to localize an unresolved target, and the amount of required frame memory A_m. For multiprocessor systems, T_{max} is the product of the maximum number of registrations with the number of time steps required to retrieve the result from the processor array. The worst case computational complexity C_{max} of a machine vision system is computed here as the product of these two parameters. Table 3.4.9-1 presents the analytical expressions for the worst case performance and computational complexity of the previously described machine vision systems. As with the average performance comparison, the systems are normalized to have the same maximum resolution (one pixel) and field-of-view (ε_0). Tables 3.4.9-2 through 3.4.9-4 present numerical values from these expressions for the fields-of-view of 512×512, 1024×1024, and 4098×4096 respectively.

Machine Vision System Type	T_{max}	A_m	C_{max}
linear foveal pattern uniprocessor	$2\varepsilon_0\log_2[\log_2(\varepsilon_0)]+2\varepsilon_0$	$2\varepsilon_0$	$4\varepsilon_0^2\log_2[\log_2(\varepsilon_0)]+4\varepsilon_0^2$
linear foveal pattern multiprocessor	$(\log_2[\log_2(\varepsilon_0)]+1)\times 2\sqrt{2\varepsilon_0}$	$2\varepsilon_0$	$(\log_2[\log_2(\varepsilon_0)]+1)\times 4\varepsilon_0\sqrt{2\varepsilon_0}$
exponential foveal pattern uniprocessor	$3\times[\log_2(\varepsilon_0)]^2$	$6\log_2\varepsilon_0$	$18\times[\log_2(\varepsilon_0)]^3$
exponential foveal pattern multiprocessor	$\log_2(\varepsilon_0)\times\sqrt{6\log_2\varepsilon_0}$	$6\log_2\varepsilon_0$	$6\times[\log_2(\varepsilon_0)]^2\times\sqrt{6\log_2\varepsilon_0}$
uniform acuity uniprocessor	ε_0^2	ε_0^2	ε_0^4
uniform acuity multiprocessor	$2\varepsilon_0$	ε_0^2	$2\varepsilon_0^3$
multiprocessor pyramid	$2\times\log_2(\varepsilon_0)$ + frame load time	$\frac{4}{3}\varepsilon_0^2$	$\frac{8}{3}\varepsilon_0^2\times\log_2(\varepsilon_0)$

Table 3.4.9-1 Analytical expressions for worst case performance and computational complexity of machine vision systems performing target localization.

Machine Vision System Type	T_{max}	A_m	C_{max}
linear foveal pattern uniprocessor	4270	1024	4.4×10^6
linear foveal pattern multiprocessor	266.9	1024	270×10^3
exponential foveal pattern uniprocessor	243	54	13×10^3
exponential foveal pattern multiprocessor	66.1	54	3.6×10^3
uniform acuity uniprocessor	131,072	262,144	69×10^9
uniform acuity multiprocessor (unfeasible)	1024	262,144	270×10^6
multiprocessor pyramid (unfeasible)	18+ set-up time	349,525	6.3×10^6

Table 3.4.9-2 Worst case performance and computational complexity of machine vision systems with a field-of-view of 512x512 performing target localization.

Machine Vision System Type	T_{max}	A_m	C_{max}
linear foveal pattern uniprocessor	8851	2048	18×10^6
linear foveal pattern multiprocessor	391.2	2048	800×10^3
exponential foveal pattern uniprocessor	300	60	18×10^3
exponential foveal pattern multiprocessor	77.5	60	4.7×10^3
uniform acuity uniprocessor	1.05×10^6	1.05×10^6	1.1×10^{12}
uniform acuity multiprocessor (unfeasible)	2048	1.05×10^6	2.2×10^9
multiprocessor pyramid (unfeasible)	20+ set-up time	1.40×10^6	28×10^6

Table 3.4.9-3 Worst case performance and computational complexity of machine vision systems with a field-of-view of 1024x1024 performing target localization.

Machine Vision System Type	T_{max}	A_m	C_{max}
linear foveal pattern uniprocessor	37.6×10^3	8192	310×10^6
linear foveal pattern multiprocessor	830.0	8192	6.8×10^6
exponential foveal pattern uniprocessor	432.0	72	31×10^3
exponential foveal pattern multiprocessor	101.8	72	7.3×10^3
uniform acuity uniprocessor	16.8×10^6	16.8×10^6	280×10^{12}
uniform acuity multiprocessor (unfeasible)	8192	16.8×10^6	140×10^9
multiprocessor pyramid (unfeasible)	24+ set-up time	22.4×10^6	540×10^6

Table 3.4.9-4 Worst case performance and computational complexity of machine vision systems with a field-of-view of 4096x4096 performing target localization.

3.4.10 Analysis of Performance Comparison

The analysis of unresolved target localization performance shows that for this task, a foveal system with exponential pattern, while requiring on average a few more foveations than a foveal system with linear pattern, requires less hardware and overall processing time. The conventional uniprocessor uniform resolution approach is woefully outperformed; its complexity measure is higher than that of the foveal systems by as much as the square of the field-of-view size, thus becoming progressively inferior at larger frame sizes (for this task). The measures C_{av} and C_{max} of the exponential system are six orders of magnitude less than the conventional array implementations and three orders of magnitude less than pyramid machines with no frame loading time.

The measure A_m indicates data structure size in the uniprocessor implementation, and both data structure size and processor count in multiprocessing implementation. Thus, A_m is a measure of system computer hardware. The small size of foveal frames translates to a significant reduction in hardware with respect to uniresolution systems implemented in either fashion.

The superior performance of foveal systems, and the exponential geometry in particular, is due to a matching of spatial resolution to the information distribution in the scene. This does not mean that the best sensor is a single pixel under saccadic control, because such an implementation cuts to a minimum system temporal resolution, and the system will have to interrogate the field-of-regard exhaustively (such a system cannot even perform saccadic gazing because there is no graded acuity in the field-of-view, or for that matter, barely any field-of-view at all).

The experiment assumed that the target was initially in the field-of-view. Without this constraint, a wider the field-of-view would enhance the system's ability to detect cues in the field-of-regard (i.e., result in higher temporal resolution). Foveal systems offer an inexpensive way to increase the field-of-view without affecting resolution at the fovea through the extension of low acuity peripheral perception. This multiresolution solution is not possible with uniform acuity systems, where spatial and temporal resolution are played against each other.

3.5 Subdivided Foveal Patterns

Figure 3.2-5 illustrates how different acuity profiles are obtained by subdividing rexels of the linear and exponential patterns. These subdivisions preserve the boundaries of the original rings, and permit the acuity within these boundaries to be scaled by any integer (Figures 3.5-1, 3.5-2). The annular region defined by the subdivision of the rexels of a linear or exponential ring is named a major ring; subdividing all the rexels of a pattern into $d \times d$ rexels forms from each original ring a major ring consisting of d rexel rings of uniform size.

Each original ring can be subdivided by a different factor, making possible a wide spectrum of acuity profiles. Indeed, any foveal geometry with square rexels and four-way symmetry can be built from combining major rings from scaled and subdivided linear patterns. For example, the exponential pattern is constructed from a linear pattern by subdividing each rexel of the linear fovea into four (2×2) rexels to form the exponential fovea, and employing scaled versions of the third linear ring (twelve rexels) as exponential rexel rings. Note that major rings from scaled and subdivided exponential patterns cannot

generate the linear pattern, e.g., no rings with odd numbered rexels along any side are possible.

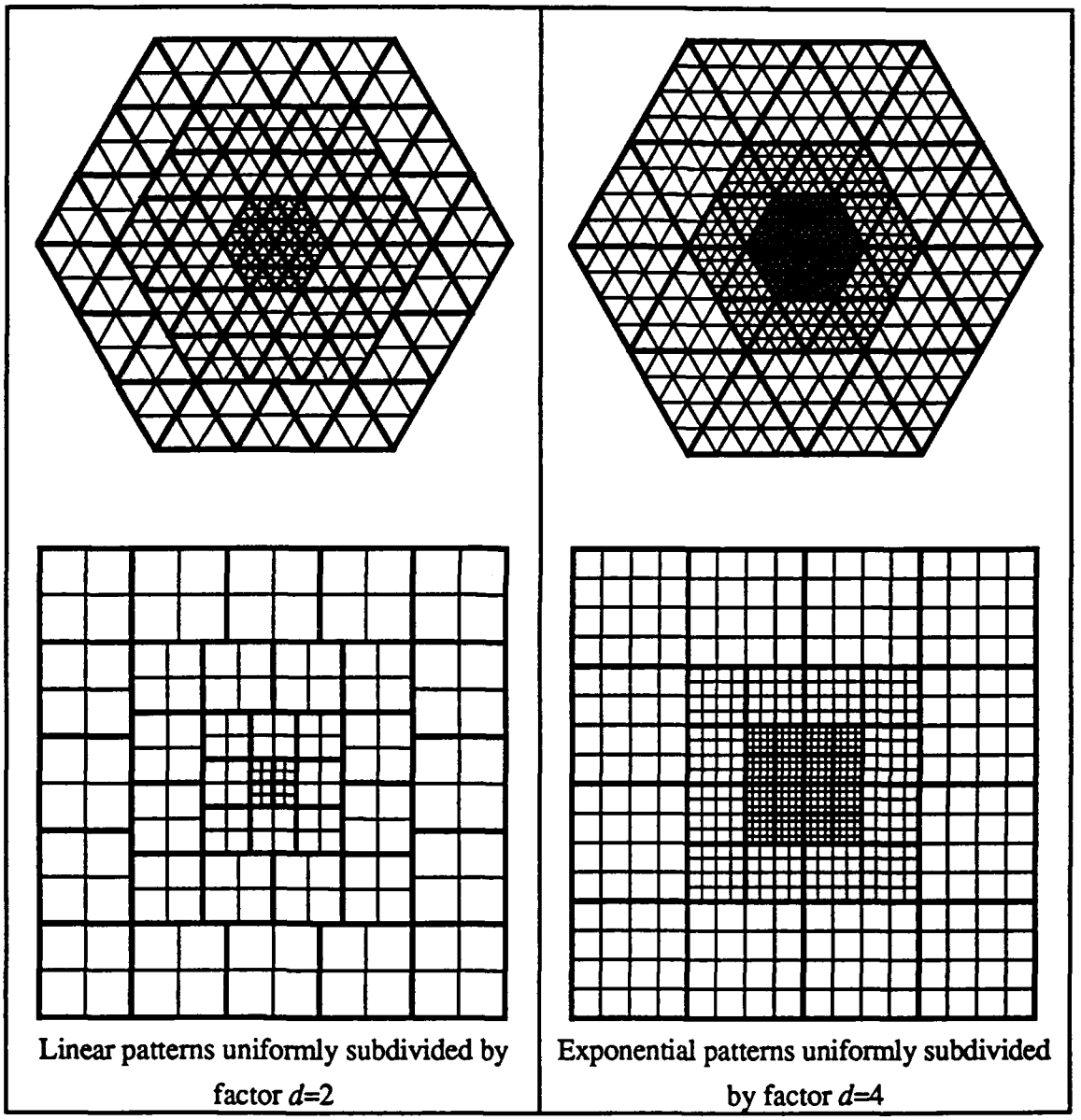


Figure 3.5-1. Examples of subdivided foveal patterns. The boundaries of major rings are highlighted, illustrating the rings and rexels of the original pattern.

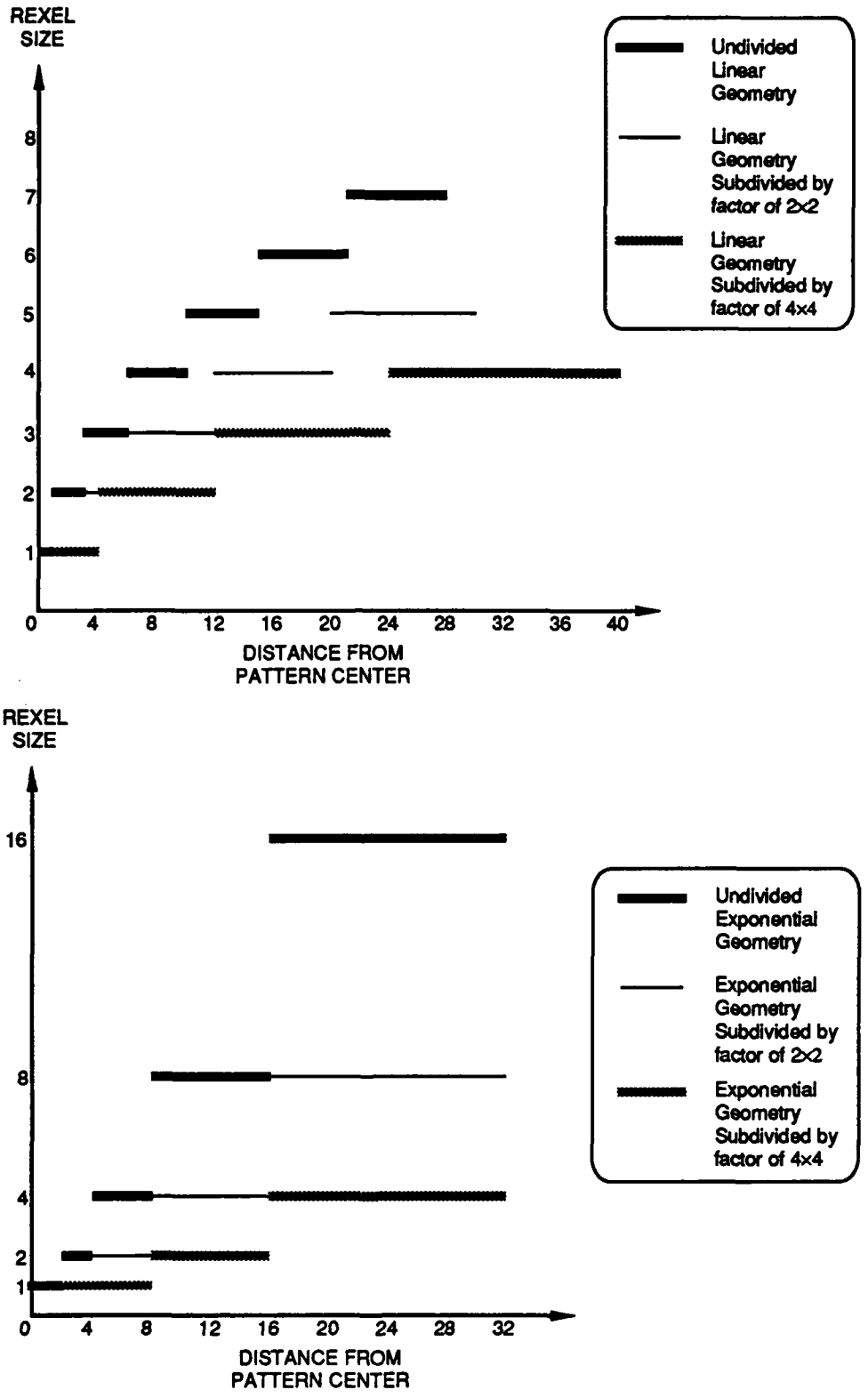


Figure 3.5-2. Acuities of subdivided foveal patterns. The rexel sizes for the original geometries and the subdivided patterns in Figure 3.5-1 are shown.

In addition to being foveal patterns themselves, the geometries generated by pattern subdivision feature regions in which rexels are uniformly sized. Within these regions, conventional space invariant signal processing techniques may be performed. Furthermore, a foveal sensor with a subdivided pattern can be piecewise constructed using conventional uniform sensor technology such as those involved in the manufacturing of focal plane array VLSI sensors.

So as to be consistent, all the subdivided patterns considered will be scaled such that the maximum acuity, i.e., the size of the smallest rexel, is that of a normalized pixel. Let $A_{r,d}$ be the number of rexels of a pattern with m rings subdivided by an integer factor d. The resulting pattern has $m'=m$ major rings each containing d rings of rexels. Since each rexel of the original geometry is subdivided into $d \times d$ rexels, the number of rexels in the subdivided geometry is obtained simply as

$$A_{r,d} = d^2 A_r \tag{3-83}$$

where A_r is the number of rexels of the original pattern, given by (3-1) or (3-12). The area of this pattern in pixels (after normalization of scale) is likewise d^2 that of the original pattern

$$A_{p,d} = d^2 A_p \tag{3-84}$$

where A_p is the area of the original pattern in pixels, given by (3-3) or (3-14). The ratio of pixels to rexels for the subdivided pattern is the same as that for the original pattern:

$$\frac{A_{p,d}}{A_{r,d}} = \frac{d^2 A_p}{d^2 A_r} = \frac{A_p}{A_r} \tag{3-85}$$

This is not to say that the reduction in foveal frame data is the same for a given field-of-view. The linear measurement of the field-of-view of the subdivided pattern is d times greater than that of the original pattern, yet the ratio A_p/A_r is a monotonically increasing function of field-of-view. Thus, (3-85) for the subdivided pattern with m' major rings is equal to that of the ratio A_p/A_r for the original pattern with m rings, but significantly inferior to the ratio of a conventional pattern with the same field-of-view.

Consider a field-of-view of size $\sqrt{A_p} \times \sqrt{A_p}$. From (3-4), an undivided linear pattern covering this space with pixel resolution at the fovea has $2\sqrt{A_p}$ rexels. The same

maximum resolution and field-of-view is obtained by subdividing a linear pattern covering the area $\sqrt{A_p d^{-1}} \times \sqrt{A_p d^{-1}}$ by a factor d, resulting in

$$A_{r,d} = d^2 2\sqrt{\frac{A_p}{d^2}} = 2d\sqrt{A_p} \tag{3-86}$$

rexels, and scaling the pattern upward by the same amount to normalize maximum resolution and field-of-view. The ratio of rexels of the uniformly subdivided linear pattern to rexels of the undivided pattern with the same field-of-view and maximum resolution is the subdivision factor itself:

$$\frac{A_{r,d}}{A_r} = \frac{2d\sqrt{A_p}}{2\sqrt{A_p}} = d \tag{3-87}$$

The range of the subdivision factor for the linear geometry is

$$1 \le d \le \frac{\sqrt{A_p}}{2} \tag{3-88}$$

At the maximum subdivision factor, the fovea encompasses the entire field-of-view and the geometry becomes a uniform lattice of pixels. Thus, the larger the subdivision factor d, the more uniform is the resulting acuity profile. This is seen in Figures 3.5-1 and 3.5-2, as rexel sizes vary less at higher d. The smallest rexel remains being the size of one pixel, but whereas the largest rexel in the undivided pattern is of linear size (3-10)

$$a_r \cong \left[A_p\right]^{\frac{1}{4}} \tag{3-89}$$

the largest rexel in the undivided pattern is

$$a_{r,d} \cong \left[\frac{A_p}{d^2}\right]^{\frac{1}{4}} = \frac{1}{\sqrt{d}} a_r \tag{3-90}$$

which is simply the number of major rings.

The same is observed for the exponential pattern. From (3-12) and (3-20), an undivided exponential pattern covering the area $\sqrt{A_p} \times \sqrt{A_p}$ with pixel resolution at the fovea has $6\log_2 A_p - 8$ rexels. The exponential pattern subdivided by a factor d with the same coverage and maximum resolution has

$$A_{r,d} = d^2\left[6\log_2\left(\frac{A_p}{d^2}\right) - 8\right] = 2d^2\left(3\log_2 A_p - 4 - 6\log_2 d\right) \tag{3-91}$$

rexels. The resulting ratio of subdivided to undivided exponential pattern rexel count is

$$\frac{A_{r,d}}{A_r} = \frac{d^2\left(6\log_2 A_p - 8 - 12\log_2 d\right)}{6\log_2 A_p - 8} = d^2\left[1 - \frac{12\log_2 d}{6\log_2 A_p - 8}\right] \underset{A_p >> d}{\cong} d^2 \tag{3-92}$$

This ratio is even more severe than with the linear pattern because the variance of rexel sizes in the exponential pattern is greater. The range of the subdivision factor for the exponential geometry is

$$1 \le d \le \frac{\sqrt{A_p}}{4} \tag{3-93}$$

At the maximum subdivision factor, the fovea encompasses the entire field-of-view and the geometry becomes a uniform lattice of pixels. The largest rexel in the undivided exponential pattern is of size

$$a_r = 2^{\frac{\log_2 A_p}{2} - 2} = \frac{2^{\log_2 \sqrt{A_p}}}{4} = \frac{\sqrt{A_p}}{4} \tag{3-94}$$

whereas that of the subdivided pattern is d times the size of the undivided pattern covering the area $\sqrt{A_p} \times \sqrt{A_p}$

$$a_{r,d} = 2^{\frac{\log_2\left(\frac{A_p}{d^2}\right)}{2} - 2} = \frac{1}{4}\sqrt{\frac{A_p}{d^2}} = \frac{1}{4d}\sqrt{A_p} \tag{3-95}$$

where, as in (3-90), the size is directed by the index of the last major ring. The difference between largest rexel size in the exponential case is a factor d, while for the linear case the difference is $\sqrt{d}$. The difference between undivided and subdivided rexel sizes for both patterns, and the fact that this difference is greater for the exponential case, is illustrated in Figure 3.5-2.

3.5.1 Acuity Profile Approximations

The exponential pattern has a number of properties which make it analytically tractable and, as shall be seen in Chapter 6, support image analysis better than the linear pattern. These properties include:

1. Scale invariance: each ring contains twelve rexels.

2. Maximum preservation of boundaries: the boundary lines between rexels in a given ring m serve as boundary lines between rexels for all rings contained therein (rings 0,1,...,m-1).

3. Power of two relationship between rexel sizes.

However, the acuity profile roll-off of the exponential pattern is very steep, and may be too severe for some applications. Specifically, peripheral acuity can be too poor to properly detect cues of importance to the vision task. For example, the acuity roll-off of the human retina is piecewise linear, with a low slope roll-off at the periphery, whereas the exponential slope increases with distance from the optical axis. Pattern subdivision is one solution to this problem. The rexels of each ring in the exponential pattern are subdivided by factors selected so that the resulting acuity profile approximates some desired profile. Unlike the uniform subdivision analyzed previously, the subdivision factor will likely vary from ring to ring.

Consider some desired acuity profile $P_a(n)$ defined at normalized integer (pixel) distances from the pattern center. Let $P_a(n)$ be indexed such that n=0,...,n_{max} is half the profile along the x or y axis of the geometry (square tessellation assumed). The error between the undivided exponential acuity profile and the desired profile is

$$e = \left(1 - \frac{1}{P_a(0)}\right)^2 + \sum_{i=0}^{m-1} \sum_{j=2^i}^{2^{i+1}-1} \left(2^i - \frac{1}{P_a(j)}\right)^2 \qquad (3\text{-}96)$$

where m is the number of rings in the exponential pattern, indexed by i along the x or y axis of the geometry (square tessellation assumed), and the first term handles the special case of the central four rexels in the exponential fovea. The error between the subdivided exponential acuity profile and the desired profile is

$$e=\left(1-\frac{1}{P_a(0)}\right)^2+\sum_{i=0}^{m'-1}\sum_{j=d2^i}^{d2^{i+1}-1}\left(2^i-\frac{1}{P_a(j)}\right)^2 \tag{3-97}$$

where m´ is the number of major rings in the subdivided geometry. The error is reduced if the rexels of each ring are subdivided by some factor d_i, where i is the ring index ranging from i=0 at the fovea to m-1 at the outer ring. It is assumed that the desired profile is monotonically non-increasing with distance from the origin. Consequently, the subdivision factor of a ring cannot subdivide the rexels to a smaller size than the size of the subdivided rexels in the preceding ring, and

$$d_i \geq 2d_{i+1} \tag{3-98}$$

$$1 \leq d_0 \leq 2^{-1}d_1 \leq 2^{-2}d_2 \leq \ldots \leq 2^{-m'}d_{m'} \tag{3-99}$$

In order to preserve the third desired property of the exponential pattern, the factors d_i must be related by an integer power of two. Thus, instead of sequence of subdivision factors d_i, the sequence of integer exponents of a scale factor of two can be employed, giving

$$d_i = d2^{\pi_i} \tag{3-100}$$

$$1 \leq 2^{\pi_0} \leq 2^{-1}2^{\pi_1} \leq 2^{-2}2^{\pi_2} \leq \ldots \leq 2^{-m'}2^{\pi_{m'}} \tag{3-101}$$

or, by taking the logarithm,

$$0 \leq \pi_0 \leq \pi_1 - 1 \leq \pi_2 - 2 \leq \ldots \leq \pi_{m'} - m' \tag{3-102}$$

The subdivision factor of the fovea $d_0=d2^{\pi_0}$ represents the scaling factor necessary to normalize the geometry to a maximum acuity of one pixel. The width of the i'th ring is thus dd_02^i pixels. From all this information, the difference between a desired acuity profile $P_a(n)$ and that of an exponential pattern subdivided by the ring associated sequence $d_i=d2^{\pi_i}$ is

$$e=\left(1-\frac{1}{P_a(0)}\right)^2+\sum_{i=0}^{m'-1}\sum_{j=d_02^i}^{d_02^{i+1}-1}\left(2^i\frac{d_0}{d_i}-\frac{1}{P_a(j)}\right)^2 \tag{3-103}$$

or, in terms of π_i,

$$e=\left(1-\frac{1}{P_a(0)}\right)^2+\sum_{i=0}^{m'-1}\sum_{j=d2^{i+\pi_0}}^{d2^{i+\pi_0+1}-1}\left(2^{i+\pi_0-\pi_i}-\frac{1}{P_a(j)}\right)^2 \tag{3-104}$$

3.5.2 Localization Performance of Subdivided Exponential Geometry

The performance of a uniformly subdivided exponential pattern performing unresolved target localization is analyzed. From (3-76) and (3-85), the expected number of foveations required to localize a target within a field-of-regard $\varepsilon_0\times\varepsilon_0$ is

$$\bar{n}_d=\frac{\log_2\left(\frac{\varepsilon_0}{d}\right)}{2+\frac{1}{3}\left[1-16\left(\frac{d}{\varepsilon_0}\right)^2\right]}\cong\frac{3}{7}\log_2\left(\frac{\varepsilon_0}{d}\right) \tag{3-105}$$

where d is the subdivision factor of the exponential geometry with the range

$$1\le d\le\frac{\varepsilon_0}{4} \tag{3-106}$$

At d=0.25ε_0, the fovea covers the entire field-of-regard and the geometry becomes uniform. It is seen from (3-105) that when d=0.25ε_0, only one registration is required to localize the target. The number of rexels in the subdivided exponential frame in terms of ε_0 is derived from (3-91) as

$$A_{r,d}=d^2\left[12\log_2\left(\frac{\varepsilon_0}{d}\right)-8\right] \tag{3-107}$$

It is seen from (3-107) that when d=0.25ε_0, the term in the brackets equals 16, and the number of rexels equals the number of pixels in the field-of-regard. The complexity measure of the system is defined as a function of d by

$$C_d=\bar{n}_dA_{r,d}=\frac{d^2\log_2\left(\frac{\varepsilon_0}{d}\right)\left[12\log_2\left(\frac{\varepsilon_0}{d}\right)-8\right]}{2+\frac{1}{3}\left[1-16\left(\frac{d}{\varepsilon_0}\right)^2\right]} \tag{3-108}$$

and represents the expected amount of data that must be processed in order to localize the target. As expected, when d=0.25ε_0, the complexity measure is simply the number of pixels in the field-of-regard. To view the relationship between complexity measure and subdivision factor, (3-109) is expressed in terms of

$$p = \frac{d}{\varepsilon_0} \tag{3-109}$$

which is continuous in the range

$$\frac{1}{\varepsilon_0} \le p \le \frac{1}{4} \tag{3-110}$$

The new complexity measure is

$$C_p = \varepsilon_0{}^2 \frac{p^2 \log_2(p)\left[12\log_2(p)+8\right]}{2+\frac{1}{3}\left[1-16p^2\right]} \tag{3-104}$$

where the argument p is proportional to d, and the first factor is a constant independent of d. The function C_p, normalized to [0,1], is given in Figure 3.5.2-1. It is seen that the subdivision factor d which minimizes the complexity measure is d=1, and that the undivided exponential foveal pattern and the uniform lattice are extremas in the complexity measure curve.

3.6 Additional Remarks

A key message conveyed by the results from this chapter is that the engineer designing a vision system should select the system's spatial and temporal resolution around the task to be performed and the expected scene. This is of course a principle consideration in the design of uniresolution systems, but the degrees of freedom (i.e., resolution and field-of-view) are limited. Foveal geometries, including their subdivided derivatives, offer greater design flexibility in which a more appropriate allocation of resolution may be selected. This selection requires an analysis of expected system dynamics, as performed here.

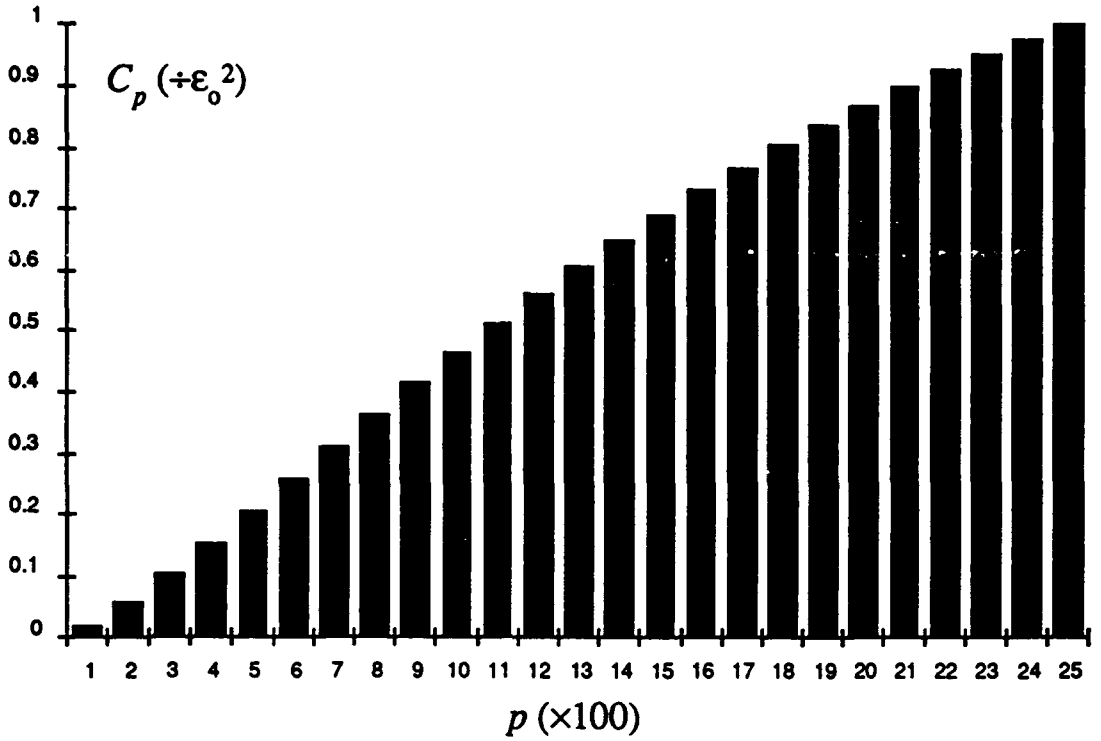


Figure 3.5.2-1. Task complexity measure as a function of subdivision factor. The argument and ordinate are normalized to be independent of field-of-view size.

The complexity measure employed in Section 3.4 did not explicitly penalize the process of moving the sensor. This overhead is difficult to measure analytically, but there are several arguments indicating it may be small. First, the mechanism for sensor steering is common to all systems, since active vision is being addressed, and no new hardware is introduced by foveal vision in this respect. The precision of the foveal steering mechanism (e.g., the number of stepper motor positions per 360° of sensor rotation) need not be at the pixel level, but roughly on the order of the fovea dimensions. Current hardware supports angular steps on the order of 0.01 degrees.

The linear and exponential foveal geometries have small foveae (2×2 and 4×4 pixels respectively) and for tasks involving resolved features, these should be larger. It is not often that the key feature of interest in a natural scene is sufficiently well resolved with such few samples. The foveae of the undivided linear and exponential geometries cover 0.0015% and 0.006% of a 512×512 field-of-view, respectively, and proportionally less for higher field-of-views. The high acuity uniform region of the human retina, on the other

hand, covers 1° or 0.5% of the total retinal area. The price paid for increasing the machine fovea to this size (instead of acuity decreasing instantly with distance from pattern center) is small since the area of the fovea still remains small. On the other hand, with this larger fovea, sensor bandwidth may be better matched to scene relevance, and no fovea (saccadic) steering precision requirements above those implemented in current systems are imposed.

The computational burden of planning saccades is small compared to frame processing. This is because saccadic planning is essentially performed by the frame processing itself. Under saccadic movement, the sensor jumps from one perceived cue of interest to another. It is the image processing algorithms working on the frame data which first identify cues and then determine what cues are relevant. When there is only one cue of interest, as in the target localization exercise, there is no saccadic planning; the image processing algorithms dictate where to look next.

When the system's vision algorithms identify several cues of interest, a gaze angle is selected which provides the maximum amount of expected relevant information. The system may decide to directly interrogate each cue sequentially, semiresolve multiple cues (e.g., foveating to the center of a cue cluster), or proceed searching for stronger cues. Chapters 5 and 6 present strategies for the selection of the gaze angle given the states of different vision tasks. While the mathematically optimal formulation of these multiple cue strategies (in the sense of maximizing expected information) can be unwieldy at times, heuristic simplifications yield computationally straightforward solutions involving minimal overhead beyond the processing requirements of the image processing algorithms working on frame data.

Servomotor control must be executed, but this is performed, in both biological and typical machine settings, by a separate processing resource which supports sensor steering from the other gaze strategies. Mechanical wear from saccades is small because the sensor remains in the field-of-view of the previous frame. Frame loading time was factored into the expressions for task execution time in the performance comparison (except in the pyramid case). The one condition where servomotor control overhead from saccades can be significant is when the system is trying to foveate onto the feature of an erratically moving target. Here, a single frame from a wide field-of-view sensor may have a higher probability of capturing the feature than the fovea. However, if the cost of the additional computational resources in the uniresolution system were invested in the mechanics of the foveal system, the latter could very well offer superior performance even in this worst case situation.

The pyramid architecture offers an attractively small latency period. In Chapter 6, pyramid architectures and algorithms are adapted to work with the exponential foveal geometry. The resulting unified foveal pyramid system exploits the low data rate and high relevant information rate of the foveal geometry to reduce conventional pyramid hardware and data structure requirements. The unified system also performs more complicated tasks faster than uniprocessor or two dimensional processor array implementations of foveal systems.

Chapter 4

Integrated Perception of Static Scenes

4.1 Introduction

The *integrated perception* is the fusion of information from a sequence of foveal sensor frames. It is thus a stable representation of the field-of-regard with spatial resolution allocated to the registration of relevant scene features. This chapter presents approaches to the generation of an integrated low level foveal perception of static scenes or dynamic scenes where integration times significantly exceed perceivable target kinematics. Just as frame data savings are obtained from the context sensitive allocation of spatial resolution within the field-of-view by the sensor, database and database processing savings are obtained by reactively allocating spatial and temporal resolution throughout the field-of-regard on the basis of information content.

A novel technique for the fusing of foveal sensor frames is presented called the discard method. This technique selectively retains frame data using acuity as a criteria. The discard method is easily implemented and generates an integrated perception which grows at a worst case rate of $O[\sqrt{n}]$, where n is the number of frames processed. As the name implies, some foveal sensor information is lost as the perception evolves. However, simulations in this chapter demonstrate that this lost information is small, and that individual sensor frames can be reconstructed from the perception with an RMS error between 0.001% and 0.1%. The reduced data of the foveal sensor frame, with respect to a uniform sensor frame with the same field-of-view and maximum resolution, and the decreasing rate of integrated perception growth with respect to number of frames processed, permits the integrated perception to represent scenes with several orders of magnitude less storage space than a single uniform sensor frame, while adequately resolving features of interest.

4.2 Perception Representation

In the vision task analyzed in Chapter 3 (target localization with the assumption of a target signature much stronger than background clutter and sensor noise), the vision system algorithm needed only the most current sensor frame. The task is sufficiently simple so as to require no memory other than a frame buffer. In more complex tasks, the integration of information from multiple frames may be necessary. Active vision in general integrates information over time into a perception knowledge base. Below are some examples.

1. The integration of multiple frames is required when the cue under interrogation is larger than the system field-of-view.

2. In a noisy environment, the system may elect to foveate to a previously visited region in order to disambiguate possible noise artifacts.

3. Object tracking obtains position information at several different times (from different frames) to predict target trajectory.

4. When tracking multiple targets distributed throughout the field-of-regard, the system foveates to targets while retaining the track data of those previously resolved but now perceived with low acuity.

In biological vision, the perception presents to higher level vision processes a representation of the environment which is more global and stable than that of a relatively small and constantly moving fovea. The perception retains the relevant details of the scene as provided by the individual retinal registrations. The perception also retains the spatial order of the relevant information. This permits a human to identify an object even though at any given time the fovea provides high resolution data on only a small area of the object. A high resolution perception of the object is obtained by foveating to key feature points on the object and integrating the feature data.

The role of perception in foveal machine vision systems is similar to its role in biological vision. The fields-of-view of different registrations can overlap, so the process of maintaining a perception in a foveal system entails the fusing of information from locations at different resolutions. Perception generation can be considered as a coding technique which transforms frame data into an integrated format. The coding can be

reversible, whereby no frame data is lost as it is integrated into the perception, or it can be irreversible (entropy reducing), whereby some frame data is lost. The advantage of entropy reducing techniques is that they yield simpler, more streamlined transforms with measurable but justifiably insignificant loss of information [Jayant84], [Lynch85].

4.2.1 Perception Levels

An integrated perception can be generated in any of the levels of the vision process. For example, the perception can consist of coded and integrated rexels from several sensor frames (low level perception) which is then analyzed for targets, or it can be the results of feature analysis on individual sensor frames (high level perception). The advantage of generating the perception at a high level is that storage space is minimized and problems of frame correlation, extremely difficult in dynamic scenes, are circumvented. However, the higher the perception level, the less information it retains. This is because the high level perception represents an interpretation of raw sensor data by models and their accompanying assumptions.

A low level perception, such as fused rexels from sensor frames of a static scene, retains all the raw data from the sensor without filtering. Another benefit of a low level perception is that it the information is usable by a greater variety of vision algorithms. For example, an integrated perception consisting of target locations supports target tracking but is of little value to target classification.

4.3 Reversible State of Nature Integrated Perception

The simplest approach to reversible data accumulation is to retain in the memory of the vision system all the sensor frames. However, besides conceptually requiring infinite memory, this accumulation approach does not generate an explicit integrated perception, and as such does not address sensor frame fusion in the true sense. Nevertheless, the accumulation approach does illustrate one important prerequisite of any reversible low level perception generation technique: any past sensor frame should be retrievable from the

perception without error. If a frame is not retrievable without error, then the low level perception generation technique failed to integrate all the frame information when it was processed, and the technique is not reversible.

The most fundamental integrated perception is the *a posteriori* probability distribution of the true state of nature (i.e., scene grey levels) given the observed data. The greater the variance of this multidimensional distribution, the greater is the ambiguity of the perception on the true state of nature. Zero variance implies that the state of nature is known exactly.[10] In actual implementations, zero variance can never be achieved because sensor noise and measurement quantization introduce ambiguity.

Let nature (i.e., the the scene) be quantized to N^2 samples (pixels), and let the value of each sample be quantized and normalized to the range of integer values $\{I \in 0, 1,..., I_{max}-1\}$. Thus each state of nature is represented by a point in I^{N^2} space, and there are $(I_{max})^{N^2}$ unique states of nature. Each state may alternately be represented by the corresponding array of N^2 pixel scalar values. The ideal low level integrated perception assigns a probability to each state. These *a posteriori* probabilities are updated after each foveation. A perception consisting of the "expected" or "most likely" state of nature retains only a moment of the overall available information. Of course, for problems of useful field-of-regard (N) and numerical accuracy (I_{max}), the size of this state space becomes astronomical. For N=512 and I_{max} =256, there are $256^{262,144} \cong 10^{631,305}$ states. This integrated perception shall nevertheless be discussed because it provides a reference for assumptions which make the implementation of integrated perceptions more feasible.

4.3.1 Algebraic Integrated Perception

The rexel data from a foveation, R, is used to update the probabilities of the states of nature. If perfect measurements are assumed (no sensor noise), every foveation reduces the number of candidate states of nature by deterministically setting the *a posteriori* probability of many states to zero. For example, if a 2×2 rexel has the value 3, then the values of the corresponding scene pixels must be one of the following 20 candidate vectors:

[10] Zero variance in the perception statistics actually indicates that the state of nature is *perceived* to be known exactly. This does not imply that nature and the perception match perfectly (i.e., they may differ by some measurement offset).

{0,0,0,3}	{0,0,3,0}	{0,3,0,0}	{3,0,0,0}
{1,1,1,0}	{1,1,0,1}	{1,0,1,1}	{0,1,1,1}
{0,0,2,1}	{0,2,0,1}	{2,0,0,1}	{0,0,1,2}
{0,2,1,0}	{2,0,1,0}	{0,1,0,2}	{0,1,2,0}
{1,0,0,2}	{1,0,2,0}	{1,2,0,0}	{2,1,0,0}

The list of candidate states of nature represents the integrated perception.[11] Data fusion is performed by filtering this list by the rexel data of consecutive frames of foveal sensor measurements. For example, let the foveal axis be redirected so that the four pixels previously measured by a 2×2 rexel with the value 3 are now measured by two 2×1 rexels (each rexel measuring two distinct pixels) with the values 1 and 2 respectively. The list of candidate states of nature for these pixels is reduced from 20 states to six: {0,1} and {1,0} for the first two pixels, and {0,2}, {1,1}, and {2,0} for the last two pixels.

This approach to data fusion is reversible. At any time, any of the frames of rexel values fused into the integrated perception can be retrieved. The simplest way of retrieving a previous frame is by foveally sampling the integrated perception itself. Any one of the states of nature in the perception, when sampled at the location of a previous foveation, produces a frame of rexel data satisfying the constraints (i.e., featuring the rexel values) against which the perception was filtered when the scene was sampled at that location.

When sensor noise is introduced, candidate states of nature cannot be rejected deterministically, and all have some *a posteriori* probability of being true. In the four pixel case, adding sensor noise would leave the number of candidate sets of values unconstrained at $(I_{max})^4$ (4.3×10^9 values for $I_{max} = 256$), each with some *a posteriori* probability derived from the model that the four scene pixels, plus noise, sum to 3. It is assumed that sensor noise can take on negative values, but not the scene signal.

This example illustrates how a rexel value can decouple the statistics of a set of pixels from the remaining pixels in the field-of-regard. Such decoupling reduces the order of the state space representation. For example, the I^{N^2} state space can be reduced after the first observation to m spaces of order I^{n_1}, I^{n_2},..., I^{n_m}, where m is the number of rexels and

[11] This assumes that scene pixel values are uniformly distributed between 0 and I_{max}-1. Otherwise, the integrated perception is represented by the list of candidate states plus their relative probabilities.

n_i, i=1,..., m is the number of scene pixels covered by the i'th rexel. Since the maximum value of n_i is on the order of $\sqrt{N}$ for the linear foveal geometry and $\frac{1}{16}N^2$ for the exponential geometry, a significant reduction in state space size is obtained. With the linear foveal geometry, the largest space size for I_{max}=256 and N=512 is reduced to 3×10^{54} states.

The statistics for all the pixels covered by a given rexel from the first frame of sensor data are the same. Consequently, a further reduction in state space size is obtained by using a single probability distribution function for all these pixels. The probability distribution of a scene pixel is determined by the information from the rexel which covers the pixel. This information consists of the value of the rexel and the rexel size. The distribution function can thus be parameterized as follows.

Let a be the value of a rexel encompassing two pixels. The number of different combinations $\beta_2(a)$ of pixel pair values $\{x_1, x_2\}$ satisfying the constraint $x_1+x_2 = a$ is obtained by counting the number of different values one pixel (e.g., x_1) can hold, while letting the value of the other pixel be determined implicitly by the constraint. Thus, x_1 can take on any value between 0 and a, and

$$\beta_2(a) = \sum_{x_1=0}^{a} 1 = a+1 \tag{4-1}$$

If the rexel is of size three pixels, then the number of different combinations $\beta_3(a)$ of pixel triplet values $\{x_1, x_2, x_3\}$ satisfying the constraint $x_1+x_2+x_3 = a$ is obtained by letting x_1 vary between 0 and a, x_2 vary between 0 and x_1, and x_3 be determined implicitly by the constraint. Thus, we have

$$\beta_3(a) = \sum_{x_1=0}^{a}\sum_{x_2=0}^{x_1} 1 = \frac{1}{2}(a+1)(a+2) \tag{4-2}$$

Similarly, the number of different combinations $\beta_4(a)$ of pixel set values $\{x_1, x_2, x_3, x_4\}$ satisfying the constraint $x_1+x_2+x_3+x_4 = a$ is obtained by letting x_1 vary between 0 and a, x_2 vary between 0 and x_1, x_3 vary between 0 and x_2, and x_4 be determined implicitly by the constraint. This gives

$$\beta_4(a) = \sum_{x_1=0}^{a}\sum_{x_2=0}^{x_1}\sum_{x_3=0}^{x_2} 1 = \frac{1}{2}(a+1)(a+2)(a+3) \tag{4-3}$$

In the general case, the number of different combinations $\beta_n(a)$ of states of nature in the region supported by a rexel of size n with value a is

$$\beta_n(a) = \frac{1}{(n-1)!}\prod_{i=1}^{n-1}(a+i) = \frac{1}{(n-1)!}\prod_{i=a+1}^{a+n-1} i = \frac{1}{(n-1)!}\frac{\prod_{i=1}^{a+n-1} i}{\prod_{i=1}^{a} i} = \frac{(a+n-1)!}{(n-1)!a!} \qquad (4\text{-}4)$$

For a rexel of size one pixel greater, (4-4) becomes

$$\beta_{n+1}(a) = \frac{1}{n!}\prod_{i=1}^{n}(a+i) = \frac{1}{n!}\prod_{i=a+1}^{a+n} i = \frac{1}{n!}\frac{\prod_{i=1}^{a+n} i}{\prod_{i=1}^{a} i} = \frac{(a+n)!}{n!a!} \qquad (4\text{-}5)$$

The number of combinations $H_{n,a}(v)$ of an n pixel set with a rexel value a and with a particular pixel x_i taking on the value v is computed by removing the pixel x_i from the set and the value v from a. Thus, $H_{n,a}(v)$ is the same as the number of combinations of an n-1 pixel set with a rexel value a-v.

$$H_{n,a}(v) = \beta_{n-1}(a-v) = \frac{(a-v+n-2)!}{(n-2)!(a-v)!} \qquad (4\text{-}6)$$

This expression assumes that the pixels are not negative, and of course, $a \geq v$. Assuming a uniformly distributed pixel value probability, the probability distribution of a scene pixel with the value v which forms a group of n pixels with an aggregate value of a is

$$P_{x_i}(v) = \frac{\text{number of pixel set combinations with } x_i = v}{\text{total number of pixel set combinations}} = \frac{H_{n,a}(v)}{\beta_n(a)} = \frac{\beta_{n-1}(a-v)}{\beta_n(a)} \qquad (4\text{-}7)$$

Only one probability distribution function per rexel is required to define the *a posteriori* probability for the state of nature after the first sensor frame of data. Unfortunately, this technique of reducing the size of the integrated perception is loses its compactness after the first sensor registration, because consecutive foveal sensor frames will group these pixel sets. Consider the rexels of frame 1 in Figure 4.3.1-1. Each rexel defines by its value and size the probability distribution of the underlying pixels. However, the shaded rexel of the next observation correlates the statistics of four rexels in the bottom right corner of frame 1. The corner rexel of frame 1 itself correlates four rexels in frame 2. In this fashion, all pixel statistics are correlated with those of all others. The

pixels within the smallest subregions defined by rexel boundaries of either frame (e.g., the four subregions within the shaded region) share the same data and thus the same statistics.[12] However, such data involves all rexels, and the regions themselves become smaller and more numerous with additional foveations, limiting their usefulness in reducing the size of the integrated perception.

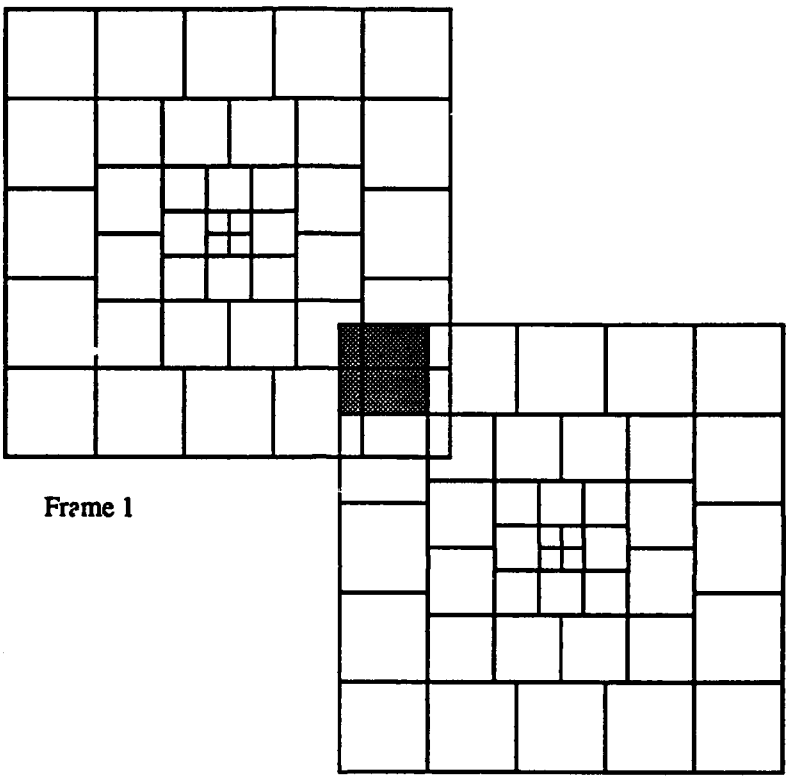


Figure 4.3.1-1. Correlation of pixel statistics through partially overlapping rexels.

The effect of partially overlapping rexels can be minimized in the case of the exponential pattern by constraining the foveal axis locations to the centers of squares in a checkerboard pattern, where the squares are the size of the largest rexel. This constraint on foveation locations ensures that rexels overlap perfectly or that groups are entirely consumed (localized correlation) as illustrated in Figure 4.3.1-2. However, this limitation permits only 16 foveal axis locations within an area equal to the field-of-regard, because that is how many of the largest rexels fit into such a region.

[12] The smallest regions defined by the superposition of the rexel boundaries of all the frames will be referred to as unisource regions.

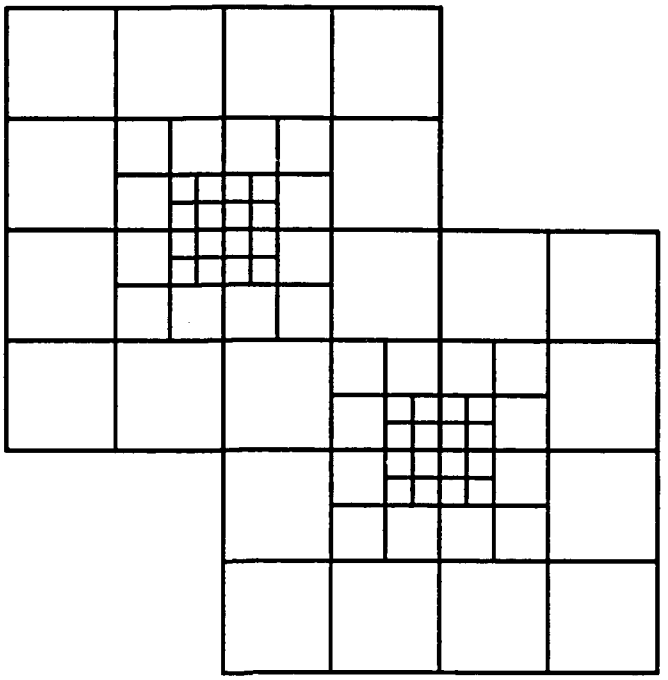


Figure 4.3.1-2. Minimal correlation of pixel statistics.

The checkerboard pattern of permissible foveations can be subdivided such that the squares are the size of the n'th largest rexel of the exponential pattern. This permits 4^{2+n} foveation locations, but also partial overlapping of rexels of larger size. Figure 4.3.1-3 illustrates how foveating to a checkerboard square the size of the second largest rexel, which provides 64 foveations per field-of-view area, can lead to correlated statistics for the larger rexels such as the one shaded.

4.3.2 Integrated Perception Using Bayesian Learning

The technique of Bayesian learning can be used to integrate the data from multiple foveal sensor frames without information loss. As opposed to the fusion of numerical values by algebraic data integration, Bayesian learning fuses the stochastic value of measurements. The *a posteriori* probability of a state of nature x_i (vector of grey levels for all scene pixels) after performing the first registration is obtained directly through Bayes rule

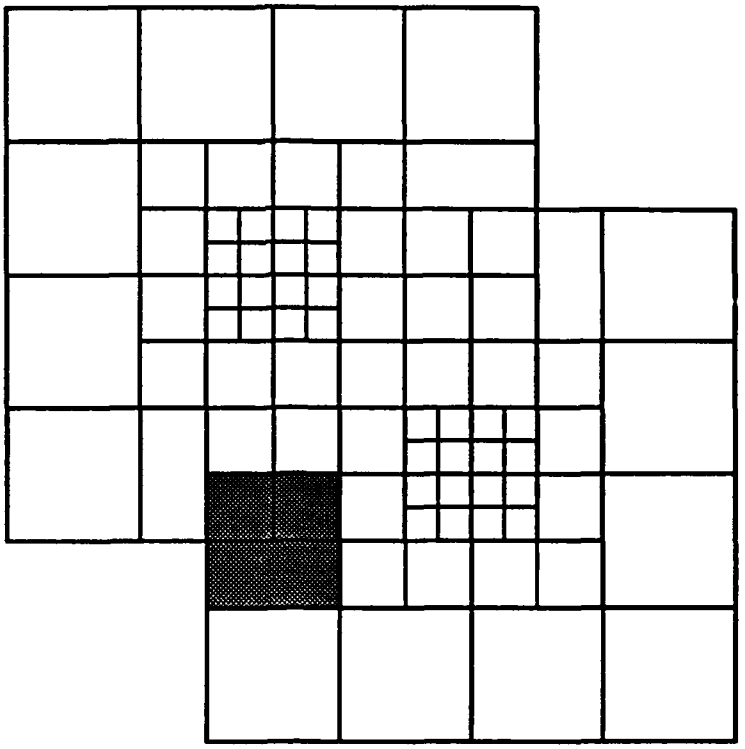


Figure 4.3.1-3. Near minimal correlation of pixel statistics.

$$P(x_i|R) = \frac{p(R|x_i)P(x_i)}{p(R)} \tag{4-8}$$

where R is the set of individual rexel data values r_j generated by the registration. The probability of the rexel data given some state of nature is expanded into

$$p(R|x_i) = p(r_1, r_2, \ldots, r_m|x_i) = p(r_1|x_i)p(r_2|x_i)\cdots p(r_m|x_i) \tag{4-9}$$

with the last term assuming uncorrelated rexels. The probability of an individual rexel value occurring given some state of nature is simply the probability of the sensor noise being the difference between the rexel value and the sum of the scene pixels encompassed by that particular rexel:

$$p(r_j|x_i) = p_n(n_{i,j}) \tag{4-10}$$

$$n_{i,j} = r_j - \sum_{\{u,v\} \in r_j} x_i(u,v) \tag{4-11}$$

The *a posteriori* probability of a state of nature x_i after performing the second registration is likewise obtained through Bayes rule with the rexel data from the first registration assigned to the "given information" field on both sides of the equation:

$$P(x_i|R_1,R_2)=\frac{p(R_2|x_i,R_1)P(x_i|R_1)}{p(R_2|R_1)} \tag{4-12}$$

Given conditionally uncorrelated measurements and substituting (4-12) for $P(x_i|R_1)$ provides the iterative expression for Bayesian learning:

$$P(x_i|R_1,R_2)=\frac{p(R_2|x_i)P(x_i|R_1)}{p(R_2|R_1)}=\frac{p(R_2|x_i)}{p(R_2|R_1)}\times\frac{p(R_1|x_i)}{p(R_1)}\times P(x_i) \tag{4-13}$$

The *a posteriori* probability after n registrations is thus

$$P(x_i|R_1,\ldots,R_n)=P(x_i)\prod_{t=1}^{n}\frac{p(R_t|x_i)}{p(R_t|R_1,\ldots,R_{t-1})} \tag{4-14}$$

$$p(R_t|x_i)=p(r_{t,1}|x_i)p(r_{t,2}|x_i)\cdots p(r_{t,m_t}|x_i)=\prod_{j=1}^{m_t}p_n\left(r_{t,j}-\sum_{\{u,v\}\in r_{t,j}}x_i(u,v)\right) \tag{4-15}$$

where m_t is the number of rexels in the t'th registration which fall within the field-of-regard, and $r_{t,i}$ is the value of the i'th rexel in the t'th registration. The inherent drawback of the Bayesian technique is that *a priori* distributions $P(x_i)$ and $P(R)$ are required.

4.4 Approximations to the State of Nature Integrated Perception

The implementation of the state of nature integrated perception is not very tractable because of the extremely large state space. This space is significantly reduced if a number of assumptions can be accepted.

The first, and probably most objectionable assumption, is that the pixels of the states of nature in the perception are statistically independent. This permits the reduction of the perception state space to just N^2 scalar probability distributions. The assumption holds

true for the first frame of rexel data, but that is not the objective of data fusion. The assumption is obviously not true when multiple measurements at different foveal axes are made, because the rexel data itself correlates their statistics. It will be shown, however, that the error introduced into the integrated perception by this assumption is small.

Given the independence of pixel statistics, each rexel value updates only the distribution of the pixels in the rexel coverage. A second assumption, which further simplifies the implementation of the integrated perception, is that the pixel distribution is Gaussian. Consider a scene of unresolved targets of value v_t against static background clutter. Let the probability distribution of the clutter pixels c_i be Gaussian with mean μ_c and variance σ_c^2. The *a priori* probability distribution of a scene pixel x is

$$f_x(x) = P(\text{target absent})\left[\frac{1}{\sigma_c\sqrt{2\pi}}e^{-\frac{(x-\mu_c)^2}{\sigma_c^2}}\right] + P(\text{target present})\left[\frac{1}{\sigma_c\sqrt{2\pi}}e^{-\frac{(x-v_t-\mu_c)^2}{\sigma_c^2}}\right] \tag{4-16}$$

Assuming a large field-of-view and relatively few targets,

$$P(\text{target present}) \cong 0 \tag{4-17}$$

$$P(\text{target absent}) \cong 1 \tag{4-18}$$

so the *a priori* scene pixel distribution can be simplified to the clutter distribution

$$f_x(x) \cong \frac{1}{\sigma_c\sqrt{2\pi}}e^{-\frac{(x-\mu_c)^2}{\sigma_c^2}} \tag{4-19}$$

Now consider a rexel of size m pixels which samples this scene. Let the value of the rexel r be the sum of the values of the m scene pixels encompassed by the rexel plus sensor noise. The noise is assumed to be Gaussian and uncorrelated with zero mean (μ_n =0). The sensor noise power will be proportional to the area of the rexel so as to represent input and thermally induced noise sources. Given the noise variance σ_{n1}^2 of a pixel sized rexel, the noise of an m pixel rexel is simply $\sigma_{nm}^2 = m\sigma_{n1}^2$.

The linear minimum mean square estimate (LMMSE) of the pixel value $\hat{x}$ given the rexel value r is

$$\hat{x} = r\left(\frac{r_{xr}\sigma_x}{\sigma_r}\right) + E\{x\} - \left(\frac{r_{xr}\sigma_x}{\sigma_r}\right)E\{r\} \tag{4-20}$$

where r_{xr} is the correlation coefficient of $f_{xr}(x,r)$, $E\{x\}=\mu_x$ is the expected scene pixel value, σ_x^2 is the pixel variance, $E\{r\}=\mu_r$ is the expected rexel value, and σ_r^2 is the rexel variance. From (4-16), $\mu_x=\mu_c$ and $\sigma_x=\sigma_c$. The resulting minimum estimate error e_m is

$$e_m = \sigma_x^2\left(1 - r_{xr}^2\right) \tag{4-21}$$

The correlation coefficient is obtained from

$$r_{xr} \equiv \frac{\sigma_{xr}}{\sigma_x\sigma_r} = \frac{E\{xr\} - \mu_x\mu_r}{\sigma_x\sigma_r} \tag{4-22}$$

where σ_{xr}^2 is the variance of $f_{xr}(x,r)$. The random variable r is composed of x plus the sensor noise n and the values of m-1 other scene pixels

$$r = x + n + x_1 + x_2 + \ldots + x_{m-1} \tag{4-23}$$

Since the components of r are assumed independent, the moments of r are simply

$$\mu_r = m\mu_x + \mu_n = m\mu_x \tag{4-24}$$

$$\sigma_r^2 = m\sigma_x^2 + \sigma_{nm}^2 = m\sigma_x^2 + m\sigma_{n1}^2 \tag{4-25}$$

Equation (4-23) can be rewritten as

$$r = x + s \tag{4-26}$$

where

$$s = n + x_1 + x_2 + \ldots + x_{m-1} \tag{4-27}$$

$$\mu_s = (m\text{-}1)\mu_x \tag{4-28}$$

$$\sigma_s^2 = (m\text{-}1)\sigma_x^2 + m\sigma_{n1}^2 \tag{4-29}$$

By being independent of x (unlike r), this new random variable s can be used to obtain r_{xr}. Specifically,

$$E\{xr\} = E\{x(x+s)\} = E\{x^2\} + E\{sx\} \tag{4-30}$$

and since s and x are independent,

$$E\{xr\} = E\{x^2\} + E\{s\}E\{x\} = E\{x^2\} + \mu_x\mu_s = \sigma_x^2 + \mu_x + \mu_x\mu_s \qquad (4\text{-}31)$$

Substituting into (4-22) gives

$$r_{xr} = \frac{\sigma_x^2 + \mu_x^2 + \mu_x\mu_s - \mu_x\mu_r}{\sigma_x\sigma_r} = \frac{\sigma_x}{\sigma_r} \qquad (4\text{-}32)$$

This gives as the LMMSE

$$\hat{x} = r\frac{\sigma_x^2}{\sigma_r^2} + \mu_x - \frac{\sigma_x^2}{\sigma_r^2}\mu_r = r\frac{\sigma_x^2}{\sigma_r^2} + \mu_c - \frac{\sigma_x^2}{\sigma_r^2}m\mu_c \qquad (4\text{-}33)$$

or

$$\hat{x} = \frac{r}{m}\Delta + (1-\Delta)\mu_c \qquad (4\text{-}34)$$

where

$$\Delta = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_{n1}^2} \qquad (4\text{-}35)$$

Note that $\frac{\Delta}{m} = r_{xr}^2$. The resulting minimum error e_m is

$$e_m = \sigma_c^2\left(1 - \frac{\Delta}{m}\right) \qquad (4\text{-}36)$$

As expected, estimate error decreases with rexel size and increases with sensor noise. The estimate error can be used as a measure of perception ambiguity at that scene pixel. In the case of no sensor noise, the pixel estimate is simply the value resulting from "spreading" the rexel measurement throughout all the encompassed pixels.

Having a convenient (first order linear) expression for the *a posteriori* perceived state of nature is part of the solution to a tractable implementation of an integrated perception. What remains is the specific algorithm implementing the fusion of rexel information from different sensor frames.

A straightforward approach is to retain only the "best" measurements on a region in the field-of-regard when data from multiple sources (different sensor frames) is made available. This approach, named the "discard method," is discussed in the following section. One may argue that this is not true data fusion, only the judicious selection of information. This may be true, and the discard method does seem to be the opposite extreme to the strict approach of maintaining an integrated perception. However, the variable acuity of foveal data provides a unique metric of the value of measurements (rexels) on which to base these judgements. This metric defines the relative information content of measurements and their value to general scene perception. Such discrimination between measurements is not possible with uniform resolution data. It will be shown that the selection of data on the basis of acuity results in the retention of almost all information (i.e., previous sensor frames are faithfully retrieved). More conventional data fusion techniques may be locally applied in borderline cases where there is little difference between the acuity of old and new information on a particular region in the field-of-regard.

4.5 Description of Discard Method

The *discard method* is a straightforward foveal data fusion method for the generation of the low level integrated perception of a static scene. This method retains the higher resolution rexels from foveal sensor frames. Rexels from a sensor frame resolving a given region of the scene are discarded when the perception already contains higher resolution data on the same region. Likewise, perception rexels are replaced by higher resolution frame rexels on the same region of the scene. The resulting integrated perception database contains the higher resolution components of overlapping foveal sensor frames. Figure 4.5-1 illustrates the perception database after the integration of three sensor frames. Each cell corresponds to a single numerical value representing the luminosity of the scene at the region subtended by the cell (i.e., a rexel datum).

The justification for the discard method is as follows. Consider M foveations, each resulting in a LMMSE $\hat{x}_i$ with error $e_{m,i}$, i=1...M, of a scene pixel p. Assuming sensor noise is independent among the measurements, the orthogonality principle gives as the optimal mean square estimate $\hat{p}$ of the pixel value

$$\hat{p} = E\{p|\hat{x}_1,\ldots,\hat{x}_M\} = c_1\hat{x}_1 + c_2\hat{x}_2 + \ldots + c_M\hat{x}_M \tag{4-37}$$

$$\sum_{i=1}^{M} c_i = 1 \tag{4-38}$$

where c_i is inversely related to $e_{m,i}$, and thus to the size of the rexel that measured the scene pixel. The discard method is an approximation to $\hat{p}$ which uses only the LMMSE $\hat{x}_i$ with the most dominant weight c_i.

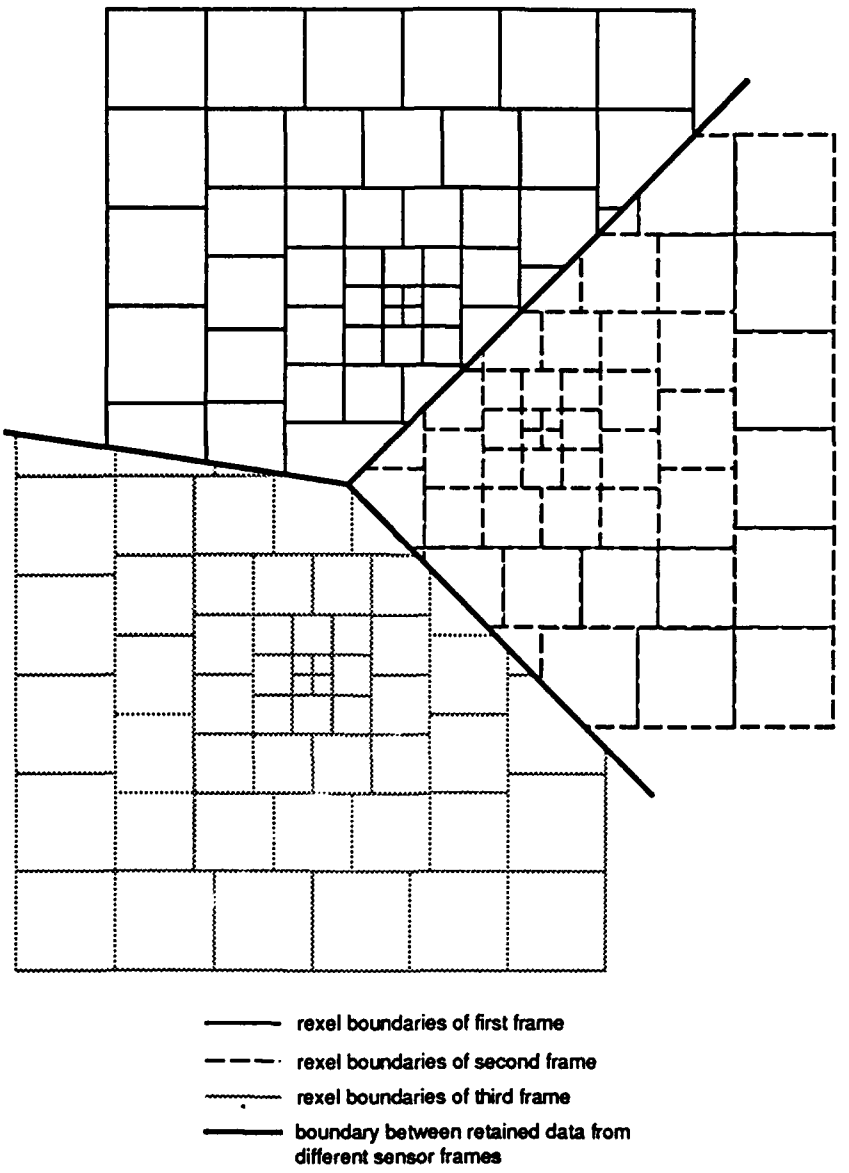


Figure 4.5-1. Integrated perception database generated by discard approach. A perception after three registrations is illustrated.

A variation to the discard approach performs a weighted average of frame and perception data (as opposed to a selection among the two). By averaging rexels, uncorrelated noise is spatially and temporally filtered out of the perception. The filtering is particularly effective in the regions where rexels of the perception and the frame being integrated are similar (e.g., the borders of equiresolution in Figure 4.5-1) and there is no particularly dominant LMMSE weight c_i. This approach requires an error variance of the perception data which provides the perception weight, and the updating of this variance every time a frame is averaged into the perception.

The discard method is not a reversible information handling process; the assumptions on pixel statistics independence artificially reduce the perception entropy, and the information discarded is not retrievable. However, it will be shown that the sensor information dropped is small because the discarded low resolution data is replaced by higher resolution data on the same scene features. No data is lost when a large rexel is replaced by a group of smaller rexels which fit perfectly within the large rexel's boundaries and there is no sensor noise.

The generation of an integrated perception by the discard approach is illustrated by the sequence of images in Figure 4.5-2. The scene being sampled is shown in Figure 4.5-2a, and consists simply of black letters on a white background (350×300=105,000 pixels).[13] The objective of the foveal system is to register the scene so as to permit unambiguous determination of the message by an optical character recognition algorithm. The optical axis is scanned twice from left to right, sampling each of the two words at three different locations. The numbered arrows indicate the foveal axis locations for the corresponding registrations. The integrated perception consists of the LMMSE estimate of the scene pixel values given the retained rexel data. The linear pattern is employed and sensor noise is assumed to be zero.

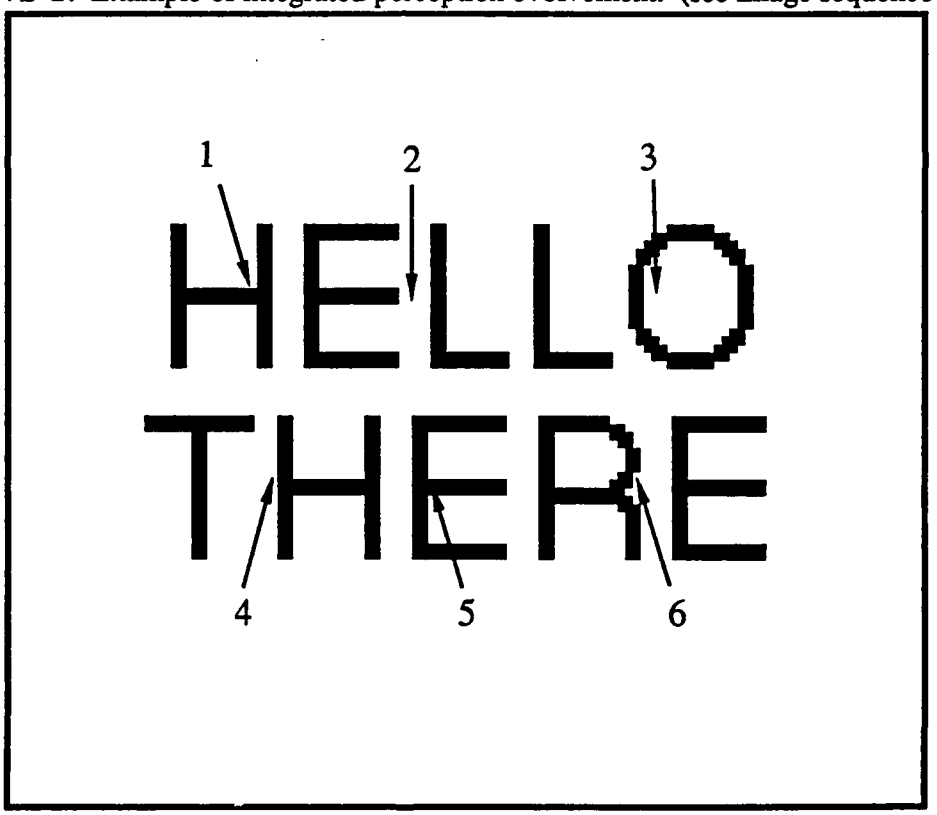
The first frame of rexel data (Figure 4.5-2b) is entirely written into the perception (Figure 4.5-2c) because there is no *a priori* information of greater (or any) acuity. A total of 647 rexels are generated. An advanced optical character recognition algorithm could correctly infer some of the letters after just this one registration, and the message might be then determined by a contextual search (albeit with some difficulty). Additional foveations

[13] This field-of-regard is smaller than most machine vision applications. Foveal systems will nevertheless offer significant advantages over uniform resolution systems, even though the advantages improve with increasing field-of-regard

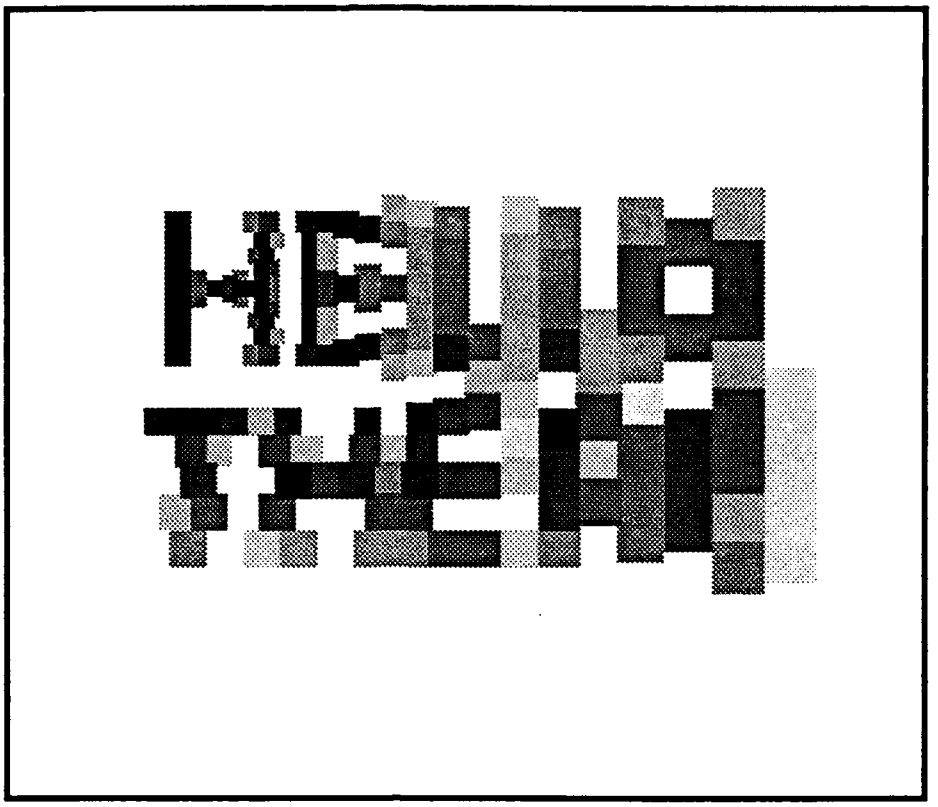
are performed to improve (reduce) the perception ambiguity and to illustrate the process of integrated perception building.

The second frame of rexel data (Figure 4.5-2d) resolves a different region of the scene. A total of 676 rexels are generated. This is greater than the number of rexels in the first frame because the optical axis of the second frame is closer to the center of the field-of-regard, and more (smaller) rexels are generated within the boundaries of the field-of-regard. The foveal system retains the pixel estimates of lesser ambiguity (expected LMMSE error), or equivalently, the smaller rexels. A total of 178 rexels from the first perception are replaced by 356 smaller rexels from the second frame. The integrated perception (Figure 4.5-2e) thus grows by 178 rexels from 647 to 825. The first word of the message can now be reliably inferred, and peripheral vision improves the perception of the second word.

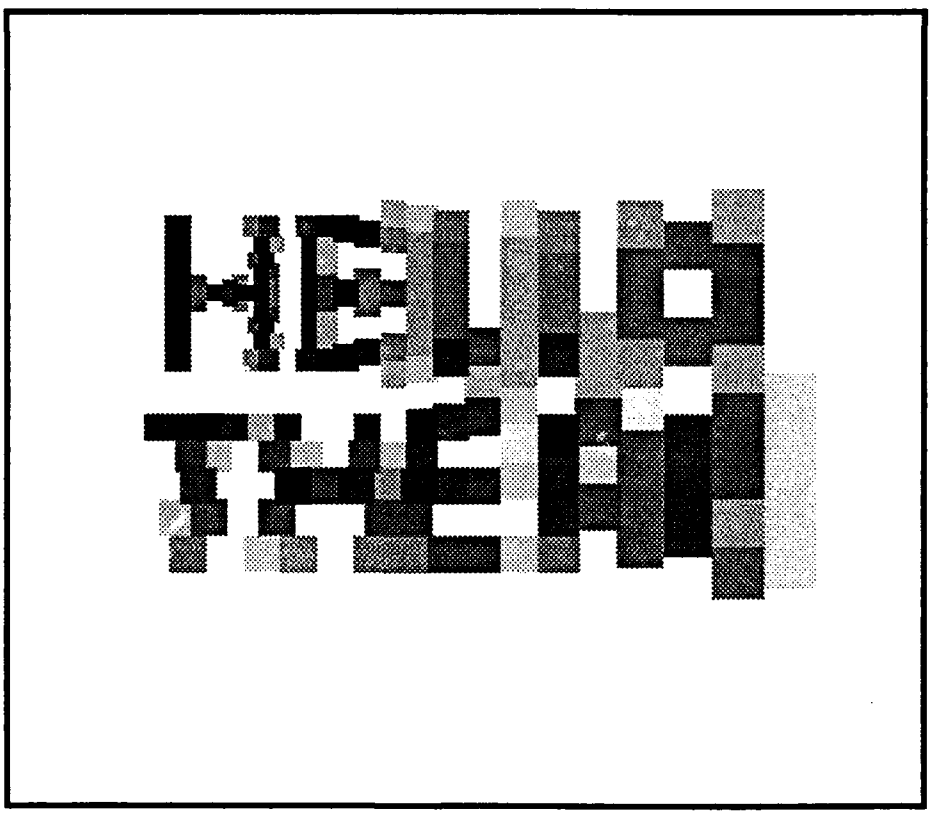
Figure 4.5-2. Example of integrated perception evolvement. (see image sequence below)



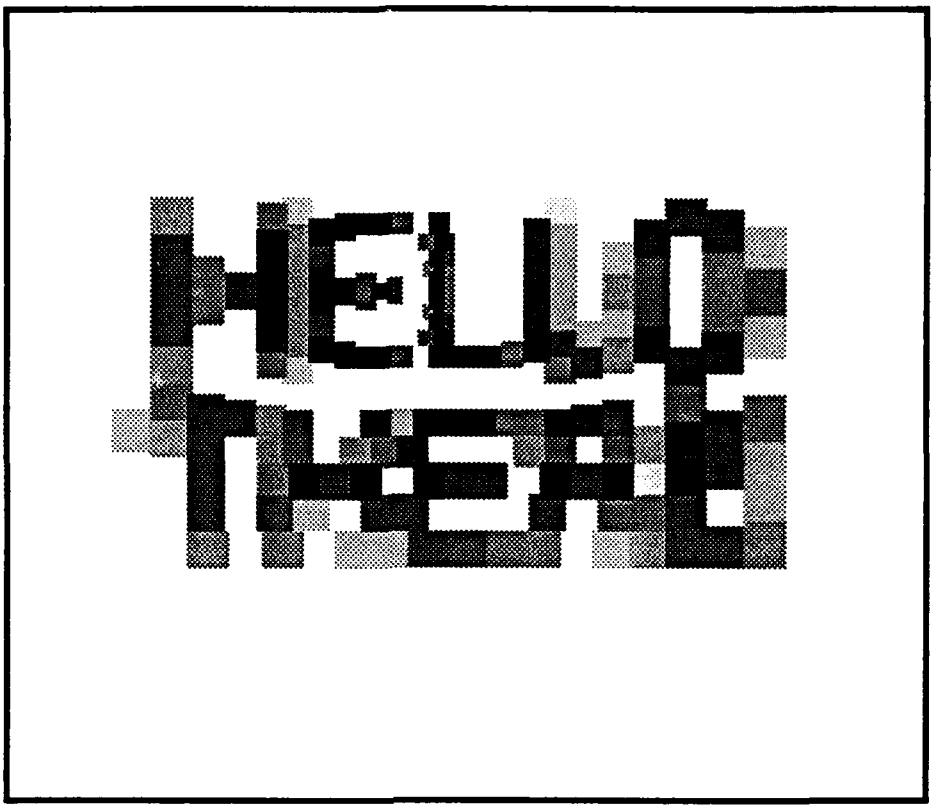
a. Test scene with foveation locations indicated.



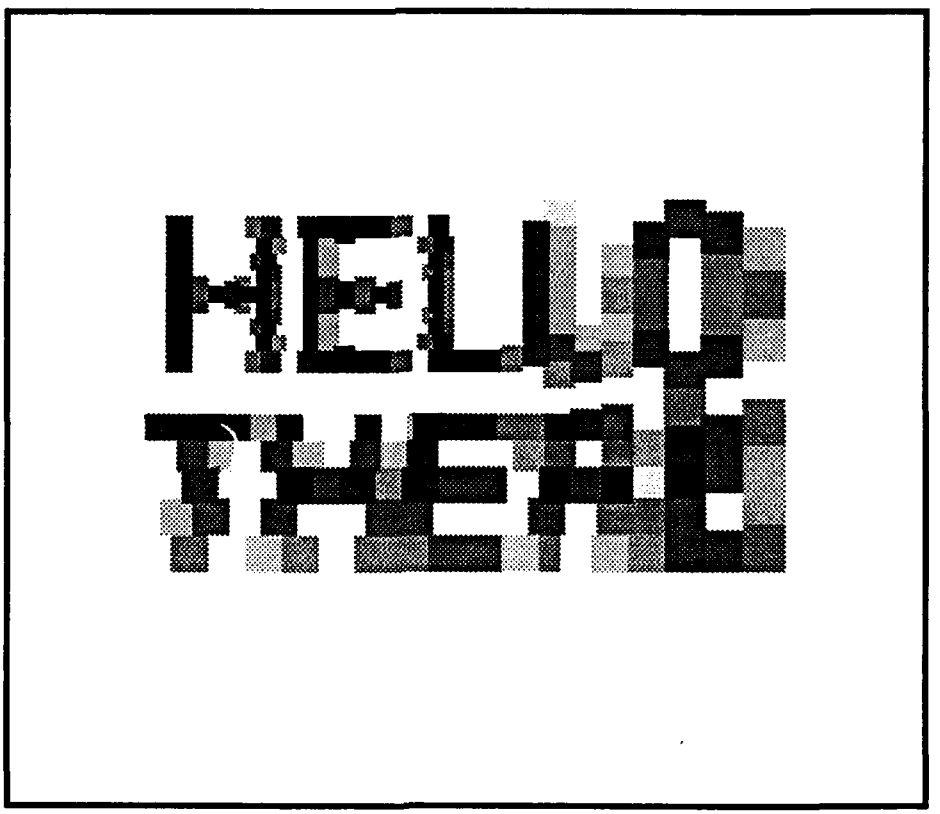
b. Sensor frame of first registration.



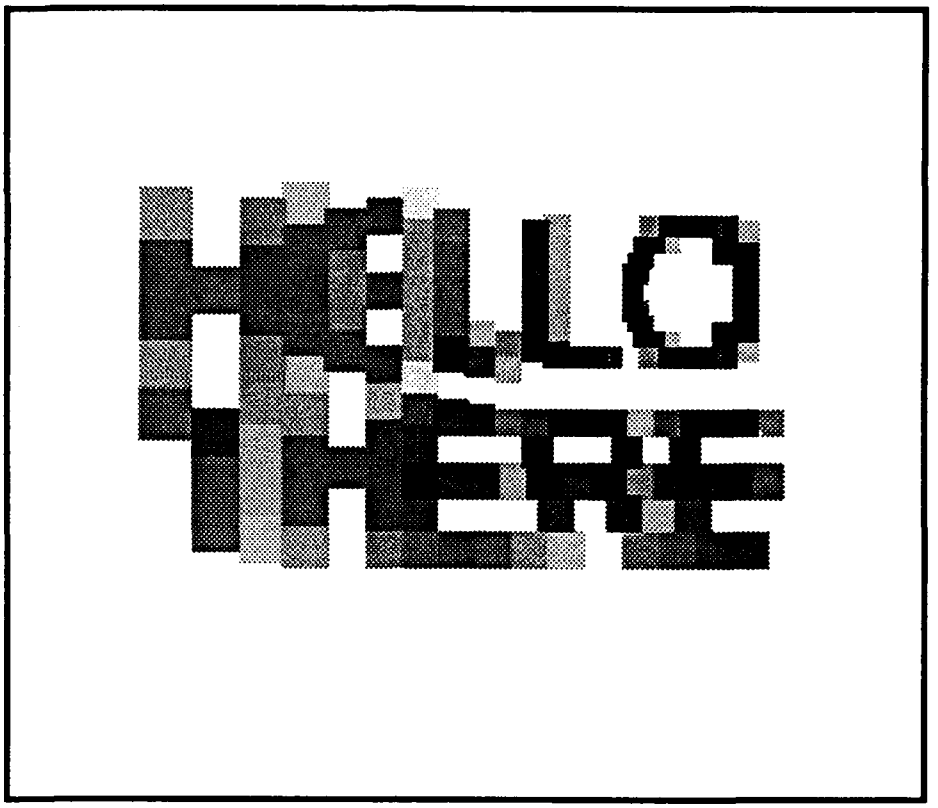
c. Integrated perception after first registration.



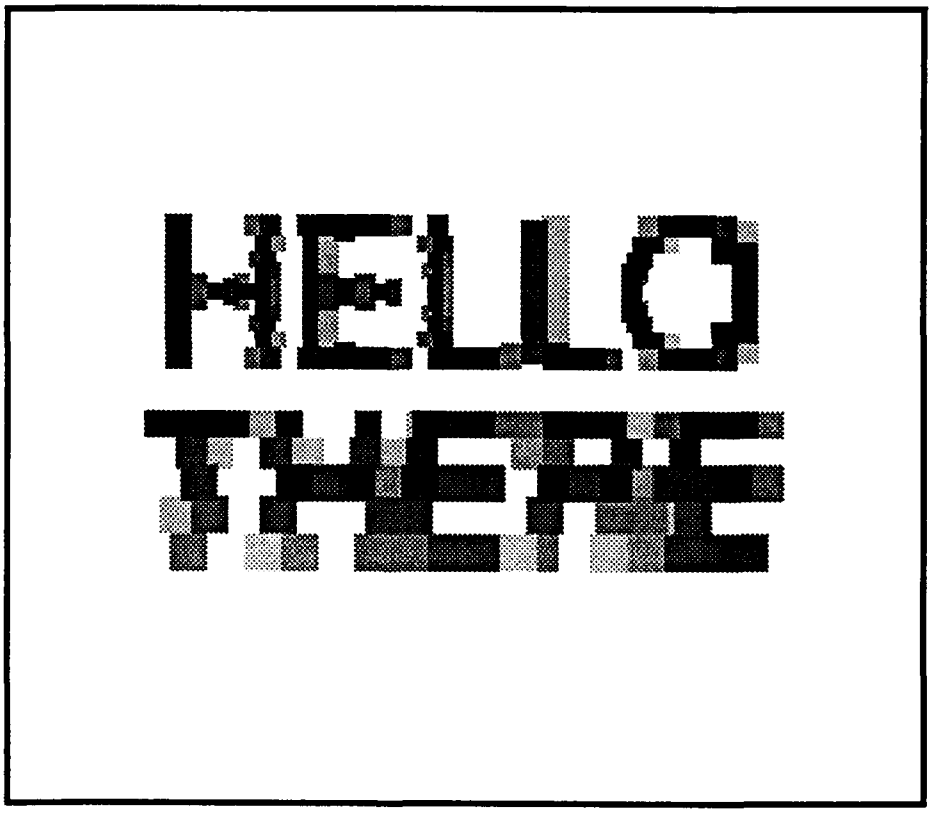
d. Sensor frame of second registration.



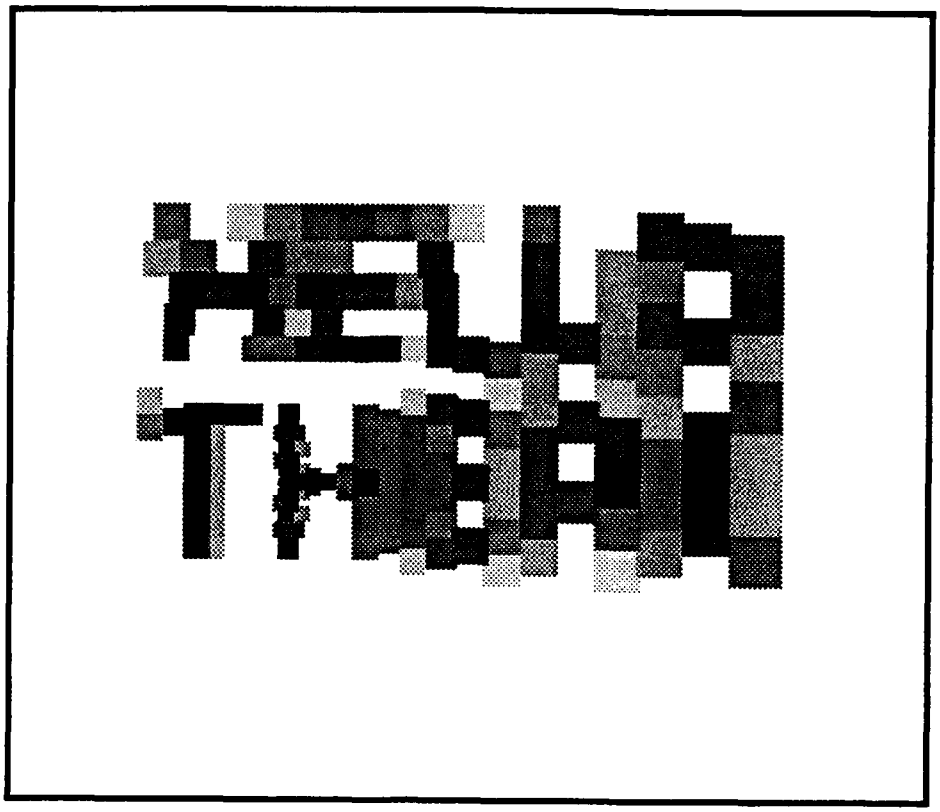
e. Integrated perception after second registration.



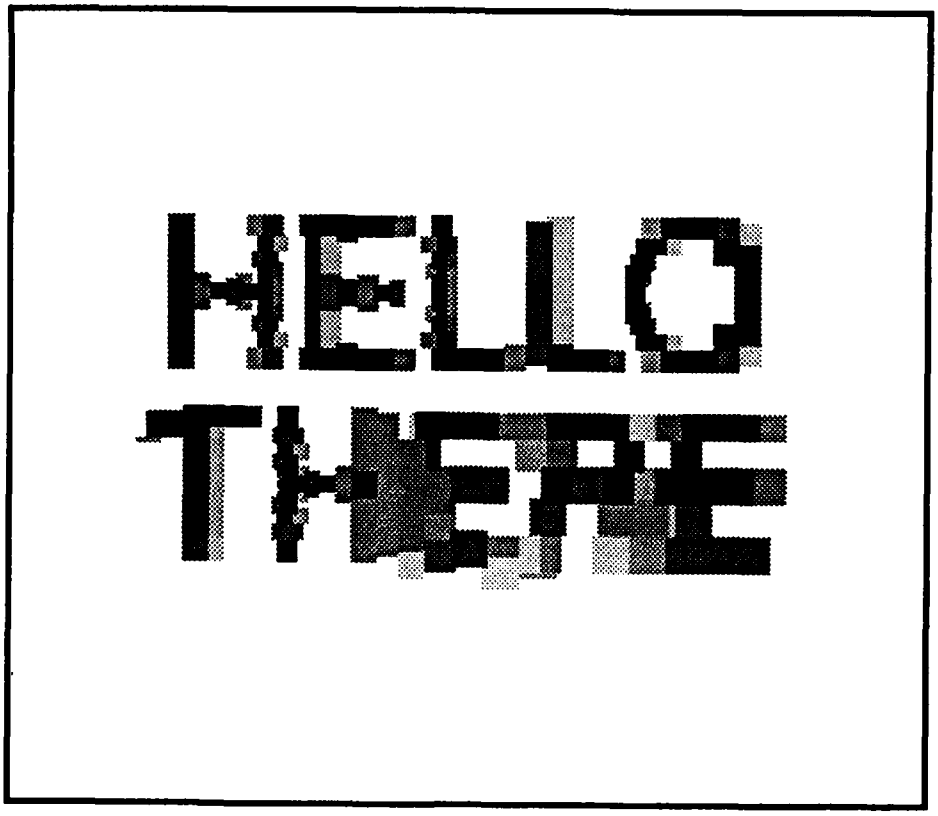
f. Sensor frame of third registration.



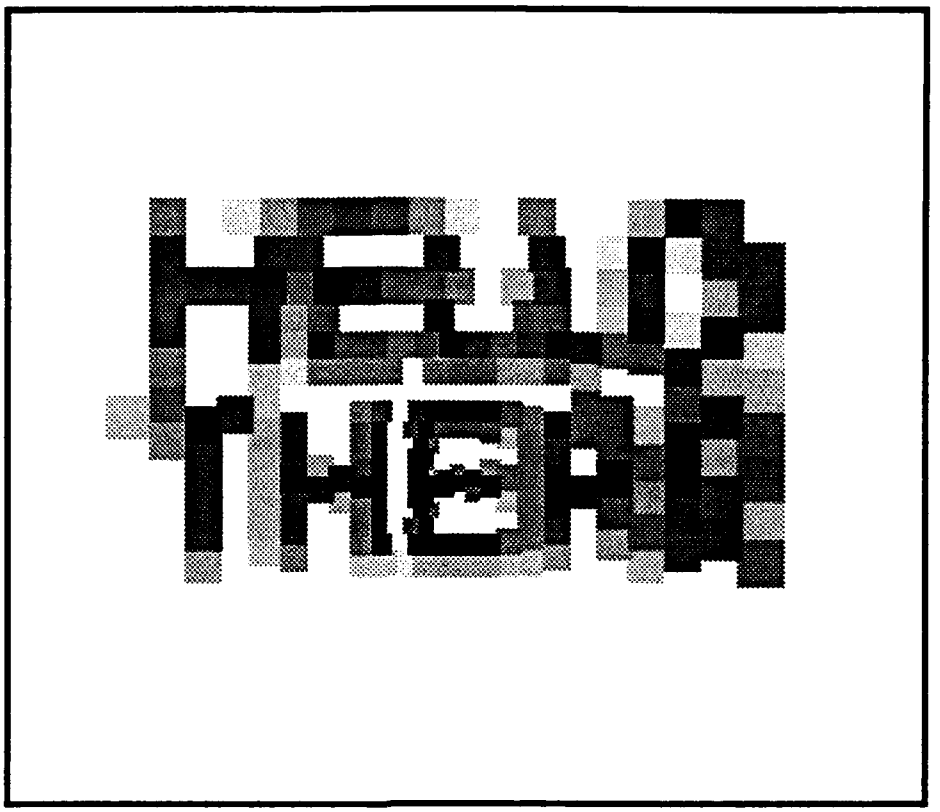
g. Integrated perception after third registration.



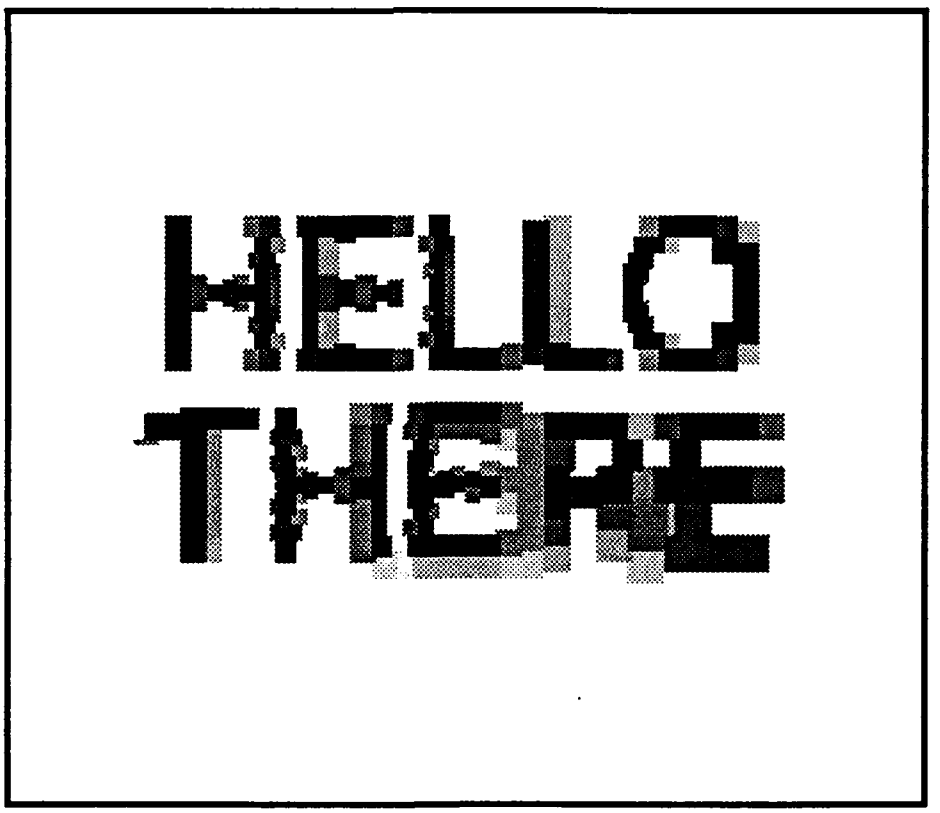
h. Sensor frame of fourth registration.



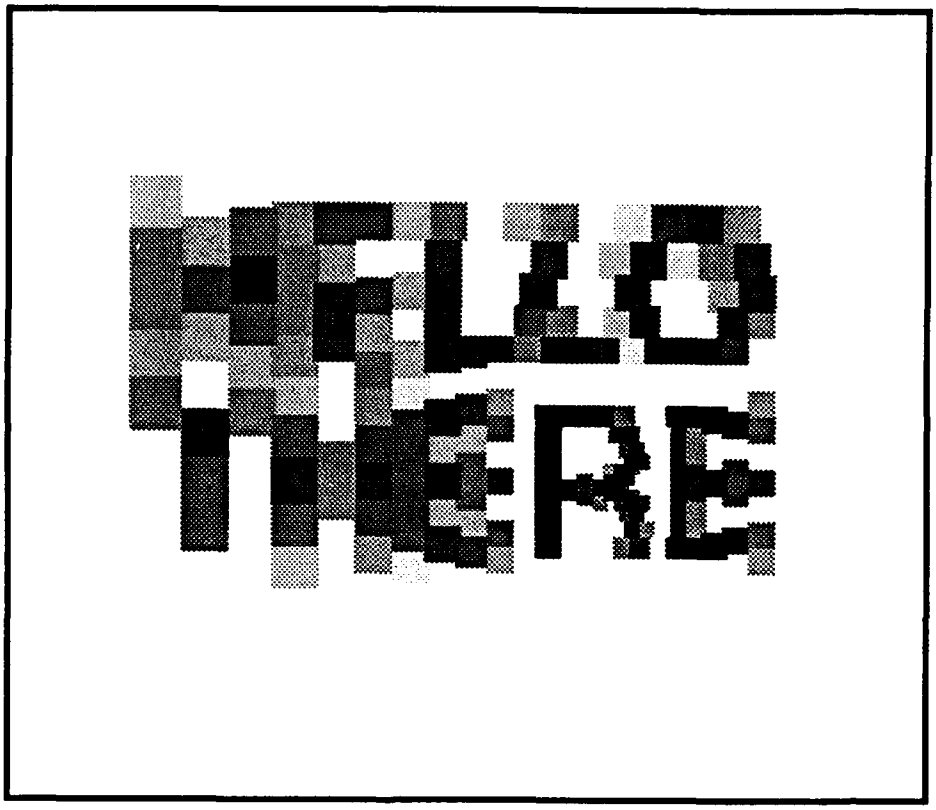
i. Integrated perception after fourth registration.



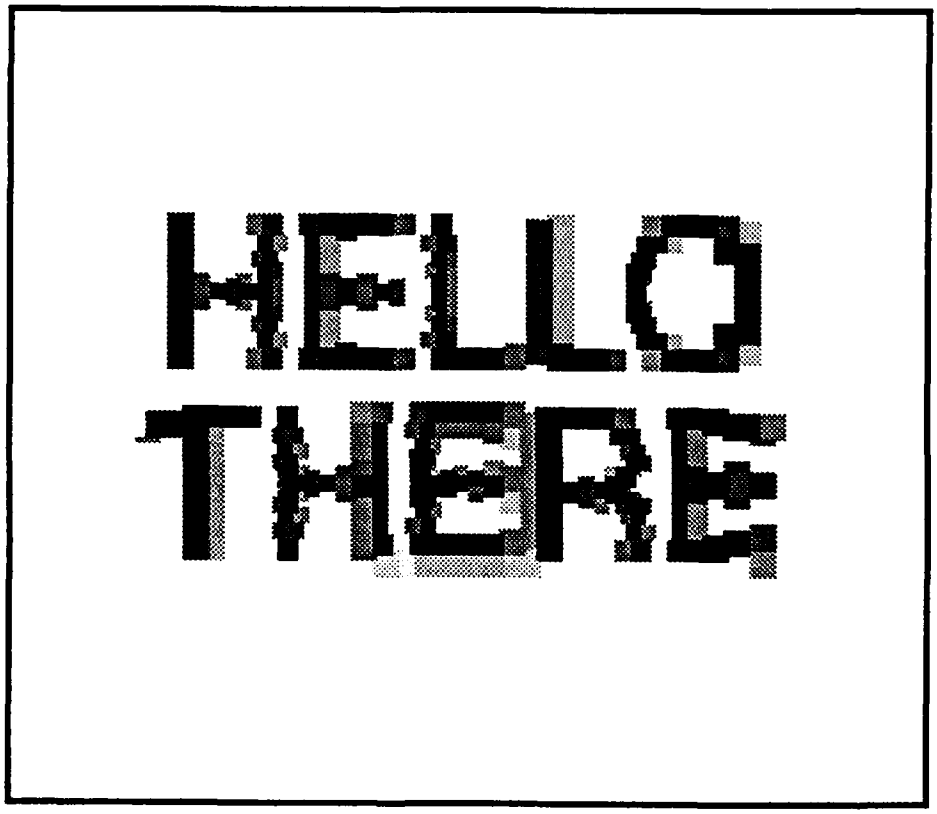
j. Sensor frame of fifth registration.



k. Integrated perception after fifth registration.



l. Sensor frame of sixth registration.



m. Integrated perception after sixth registration.

The third registration generates 666 rexels (Figure 4.5-2f). A total of 126 rexels from the current perception are replaced by 309 smaller rexels from the third frame. The integrated perception (Figure 4.5-2g) thus grows by 183 rexels from 825 to 1008. The first word of the message is now well resolved, and second word can be reliably inferred.

The fourth, fifth, and sixth registrations (Figures 4.5-2h, j, l) produce a perception (Figures 4.5-2i, k, m) that resolves the second word of the message. The perception of the first word remains unchanged during these foveations because the foveal axis of these last three registrations are farther from the word than the axis of the first three registrations. Consequently, no information on the first word is received from the last three frames with greater acuity than that previously received and integrated.

The number of rexels obtained and retained from each registration, and the number of discarded perception rexels, is presented in Table 4.5-1. It is seen that the rate of growth of the integrated perception generally decreases with the number of foveations. This is to be expected, since the scene is static and as the perception acuity improves, the "appeal" of sensor data diminishes.[14] The total number of rexels in the integrated perception is 1424, a significant reduction from 105,000 (pixel) values. The savings are obtained primarily by the low resolution sampling of the irrelevant white border around the text.

Registration	Rexels in sensor frame (in FOR)	Rexels dropped from perception	Frame rexels added to perception	Overall growth in perception	Size of perception in rexels
1	647	0	647	647	647
2	676	178	356	178	825
3	666	126	309	183	1008
4	650	136	310	174	1182
5	696	96	208	112	1294
6	668	99	229	130	1424

Table 4.5-1. Data retention and discard in "Hello There" example of integrated perception evolvement.

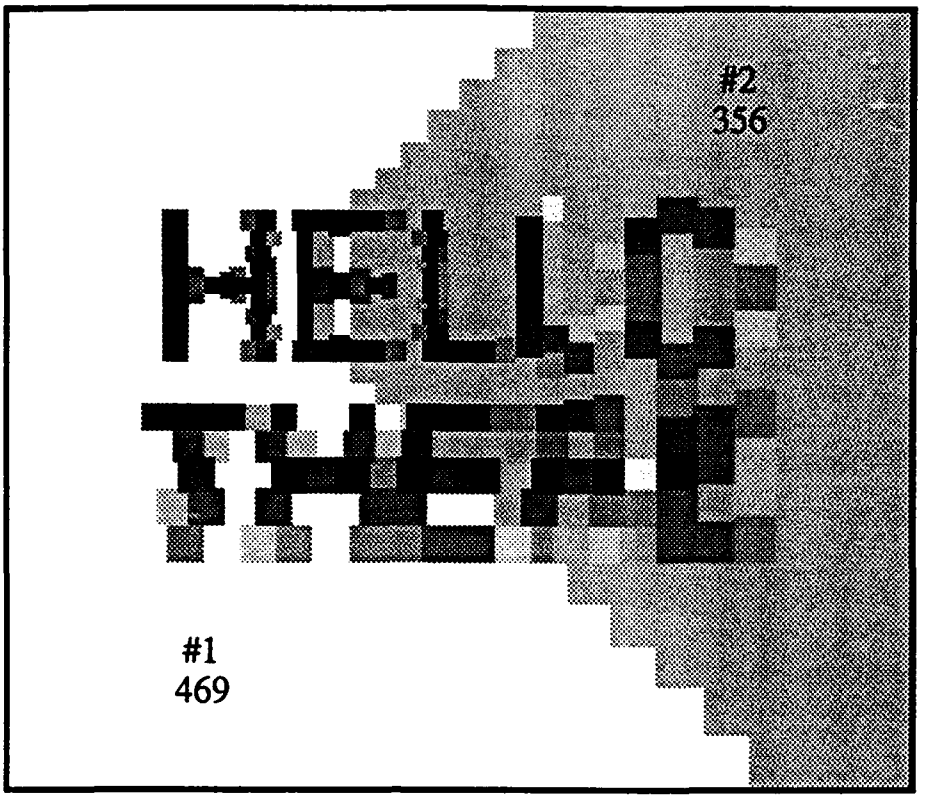
[14] In dynamic scenes (e.g., moving objects, moving system platform), all frames are important to the vision system, conveying information on new scene states and objects/features not present in earlier registrations (e.g., hidden surfaces).

Figures 4.5-3 illustrates the information each sensor frame contributes to the integrated perception. The perception is segmented into regions, each containing data contributed by a single frame. Each region is tagged by the frame number which contributed the data and the approximate number of rexels in the region. The shape of the regions is determined by the placement of the foveal axis and by the data fusion criteria of retaining frame rexels of greater acuity than those present in the perception. If the criteria were to retain frame rexels of greater *or equal* acuity, the regions would different (Figure 4.5-4).[15] However, for static scenes and noise-free sensors, the integrated perception information is the same.

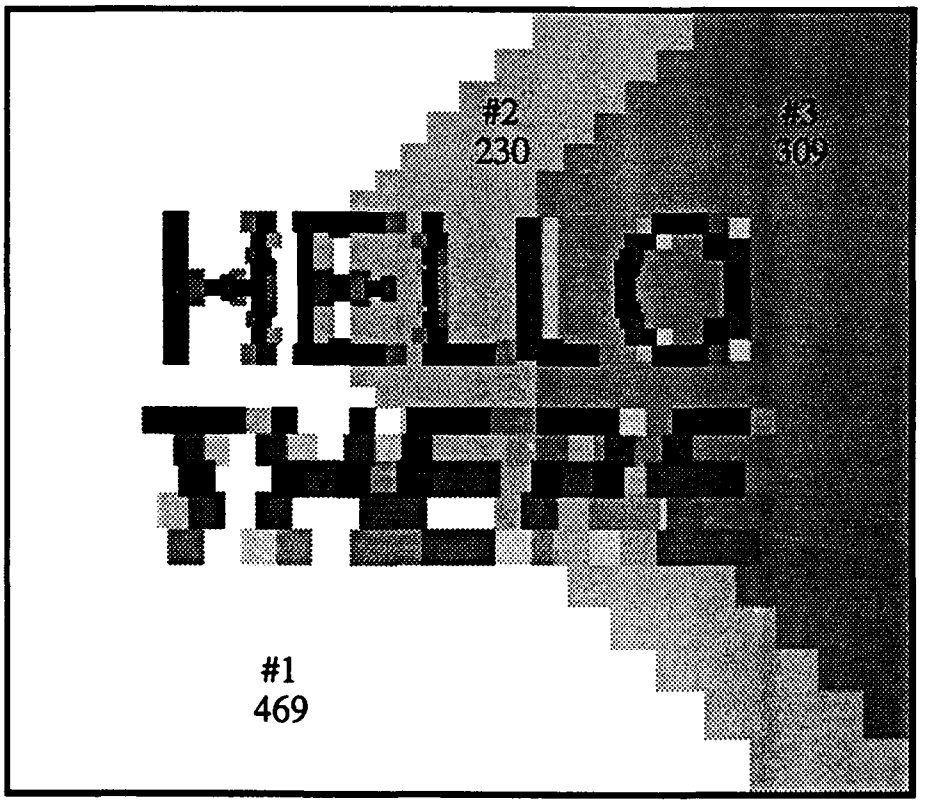
Figure 4.5-3. Distribution of perception data in example of integrated perception evolvement. The data retained from each frame is labeled by the frame number within the foveation sequence and the number of rexels in the data (see image sequence below).

a. Integrated perception region after first registration.

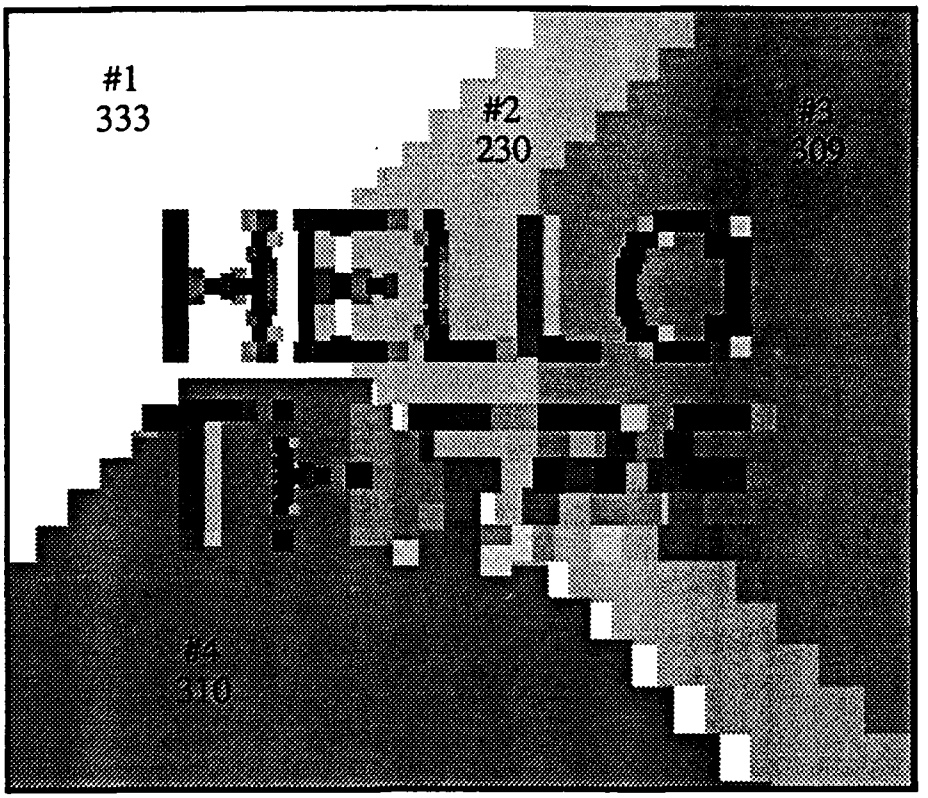
[15] The ribbed edge effects visible at region boundaries in Figure 4.5-4 and region 5 in Figure 4.5-3 appear when the foveal axes are skewed, and alternating slices of rexel rings from different frames offer the greater acuity.



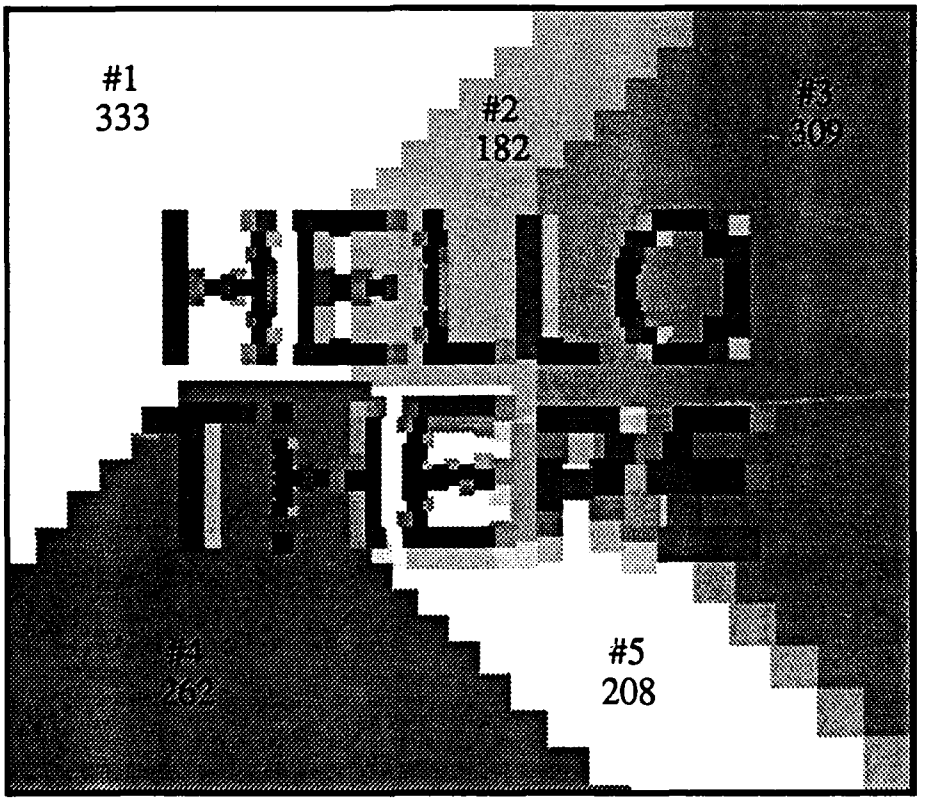
b. Integrated perception region after second registration.



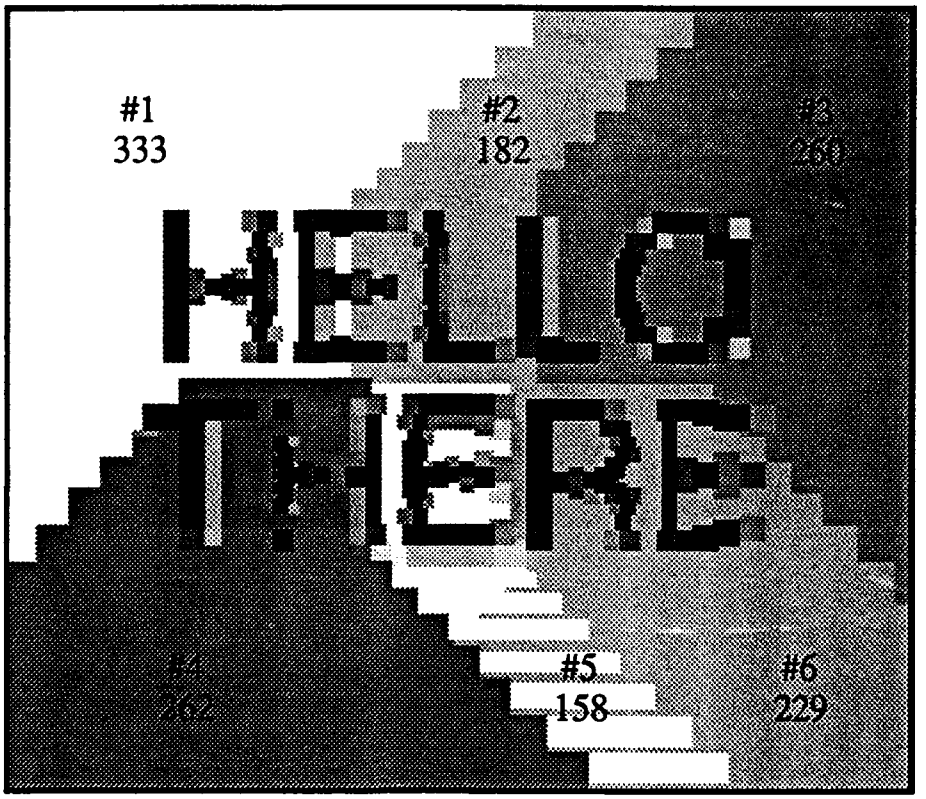
c. Integrated perception region after third registration.



d. Integrated perception region after fourth registration.



e. Integrated perception region after fifth registration.



f. Integrated perception region after sixth registration.

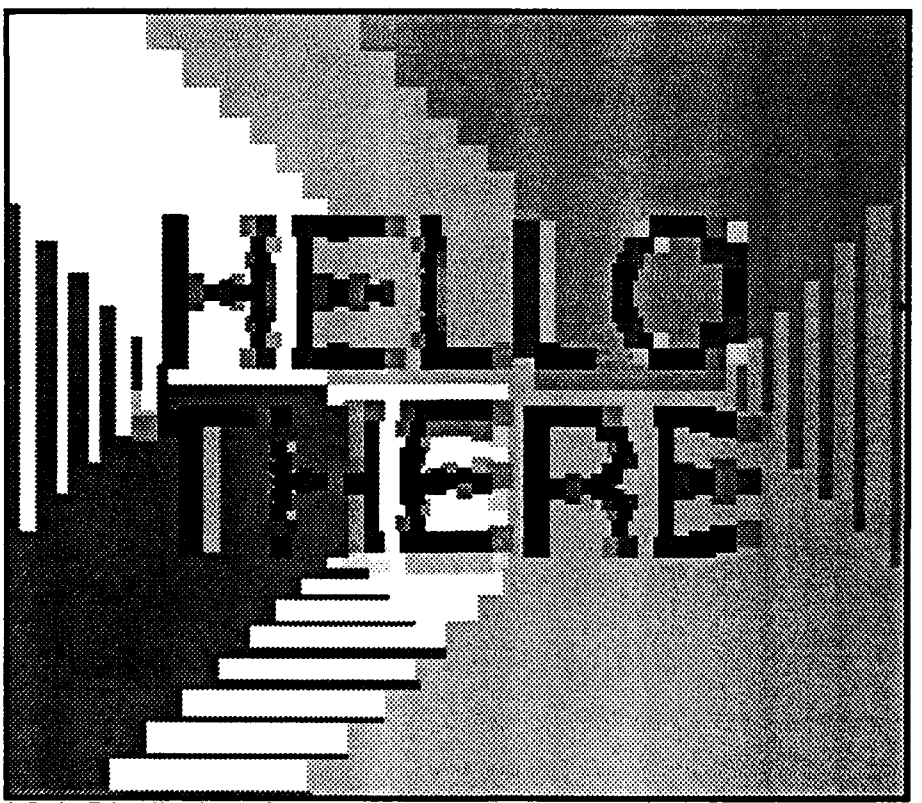


Figure 4.5-4. Distribution of perception data after sixth registration using greater than or equal acuity retention criteria.

4.6 Acuity Profile of the Integrated Perception

The foveal integrated perception consists of two fundamental types of information: the measured scene luminosity, and the resolution (acuity) with which each measurement was made. The former is used to support the scene understanding functions of the machine vision system, just as with uniform resolution machine vision systems. The latter, which has no counterpart in uniform resolution systems, is necessary to select the frame data to be retained from each foveation, and properly interpret the measurement data. For example, a rexel measuring 1×1 pixels and another rexel with the same value measuring 10×10 pixels have very different implications.

The acuity of the integrated perception generated by the discard method is a superposition of the sensor acuity profile centered at the different foveal axis locations. The acuity profile for the linear and exponential foveal sensors is illustrated in Figures 4.6-1 and 4.6-2, respectively, for a field-of-view of 512×512 pixels (the x-y plane represents the focal plane and the z axis represents acuity). This function is proportional to the inverse of the linear dimensions of the rexels. The square of the acuity profile can be interpreted as a measure of localized data density, or the proportion of rexel area to unit area.

Figure 4.6-3 illustrates the acuity profile of the integrated perception formed by the six registrations in the example of Section 4.5. The x-y plane in this case represents the system's field-of-regard. Since only the highest acuity data is retained by the discard method, the acuity profile of the integrated perception peaks at the locations of the x-y plane corresponding to the locations in the field-of-regard where the foveal axis was directed. The acuity profile of the perception after the integration of each frame in the example of Section 4.5 is illustrated in Figure 4.6-4. The brighter the image, the higher the acuity. The banding of acuity caused by the discrete steps in rexel size is clearly visible.

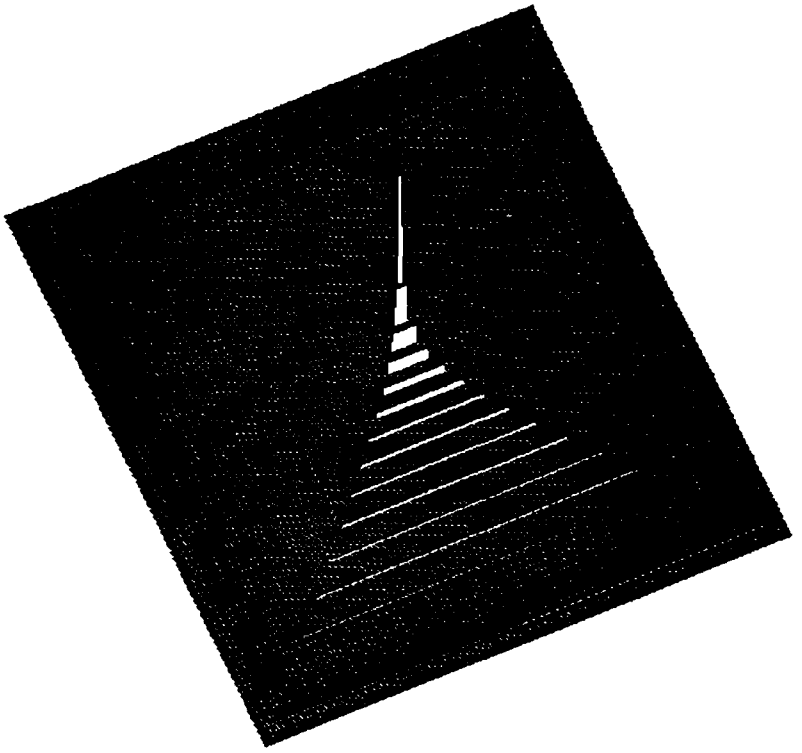


Figure 4.6-1. Acuity profile of the linear geometry (512×512 pixels).

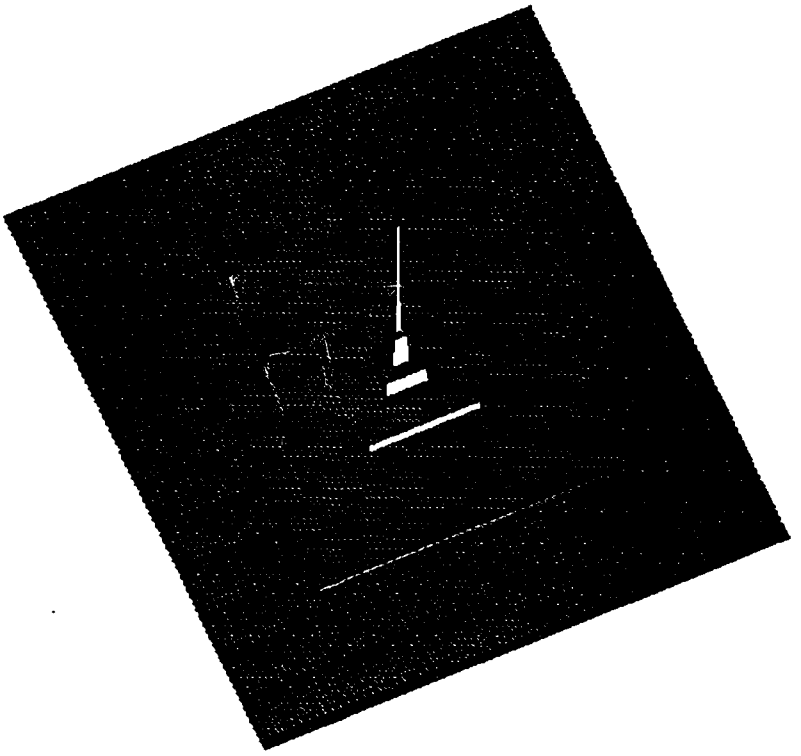


Figure 4.6-2. Acuity profile of the exponential foveal geometry (512×512 pixels).

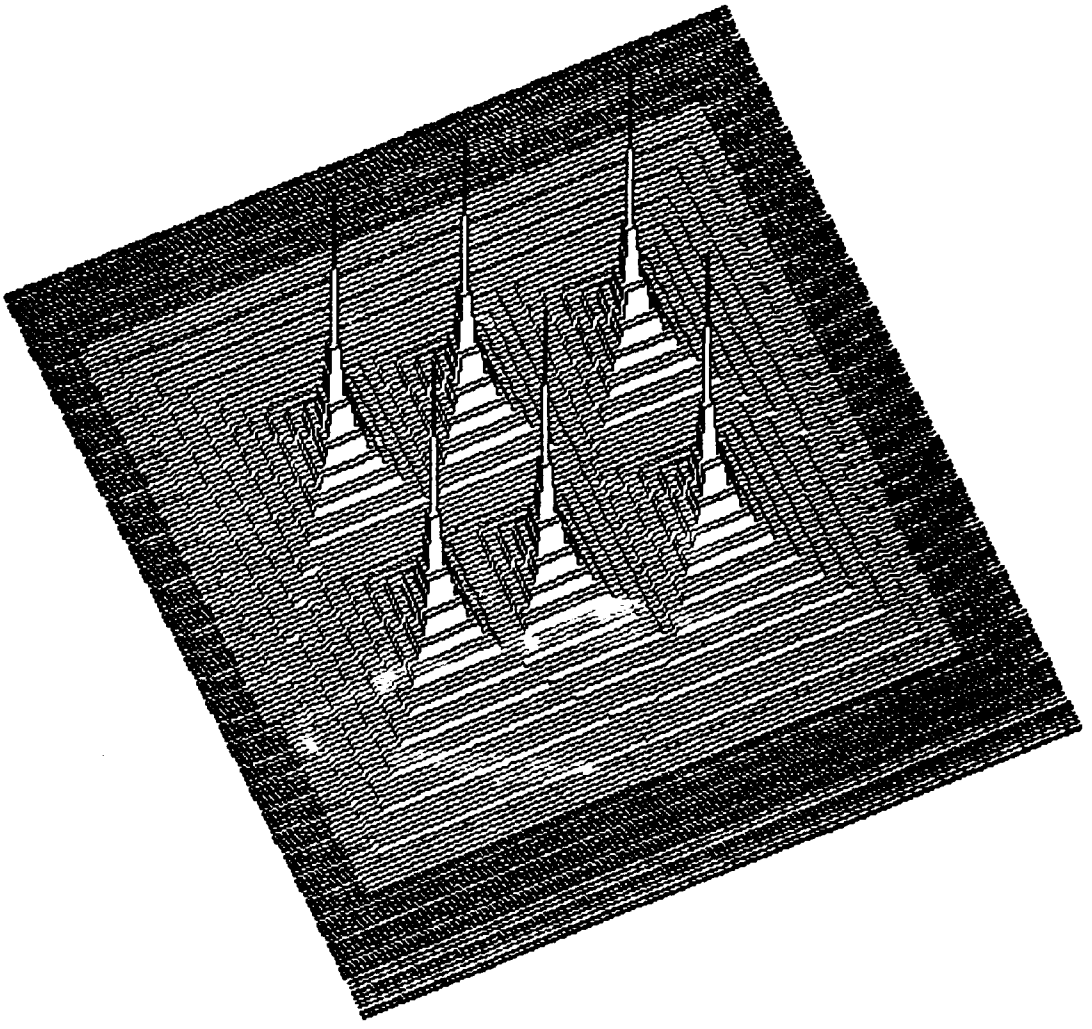
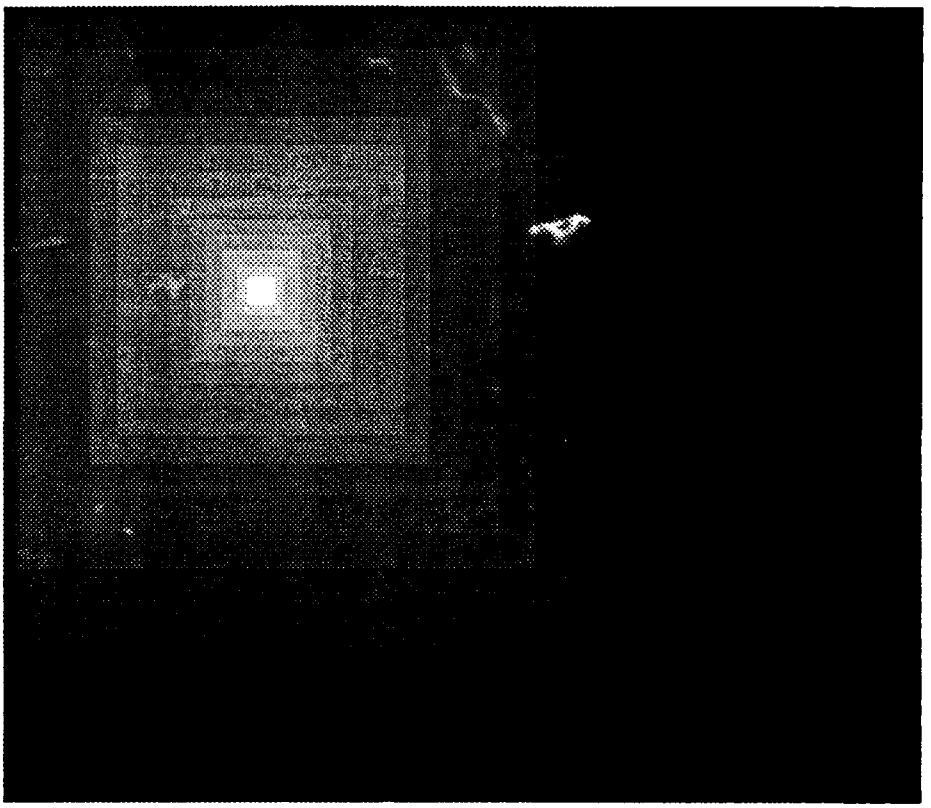
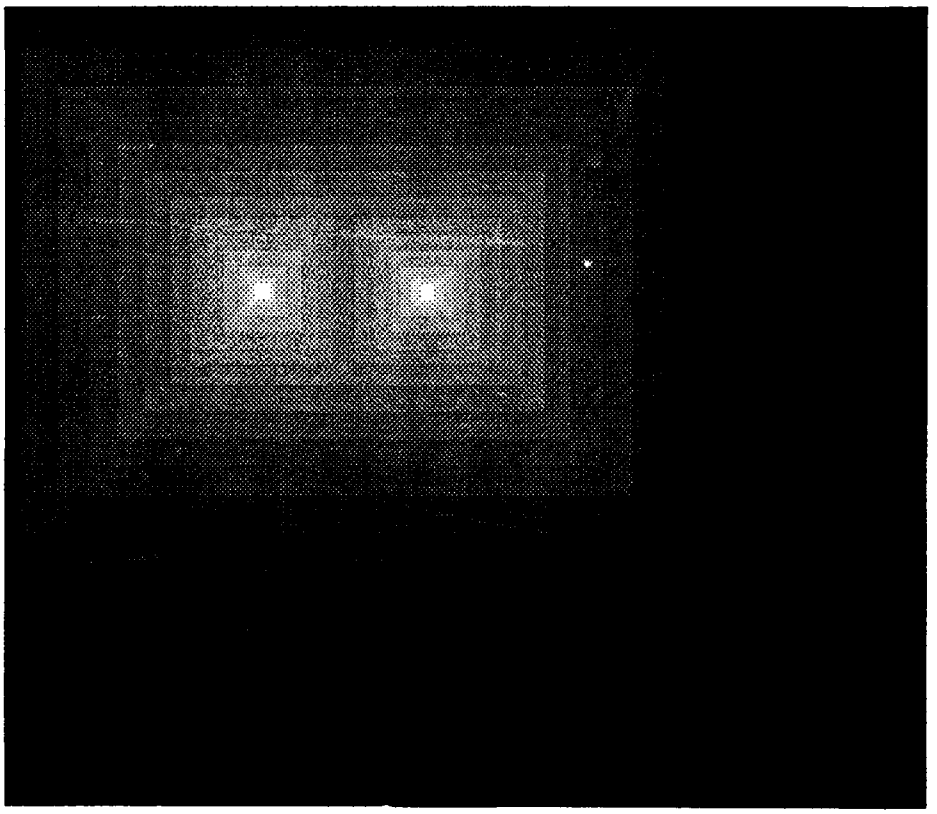


Figure 4.6-3. Acuity profile of the final integrated perception in the "Hello There" sequence.

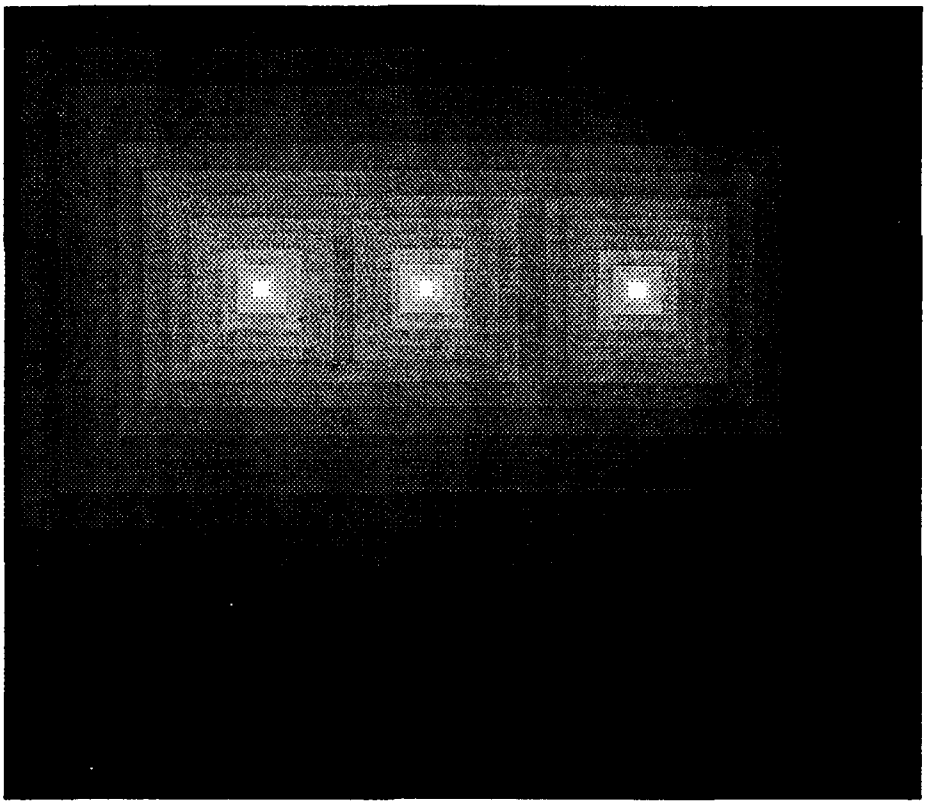
Figure 4.6-4. Acuity profile of integrated perception. (see image sequence in the following pages)



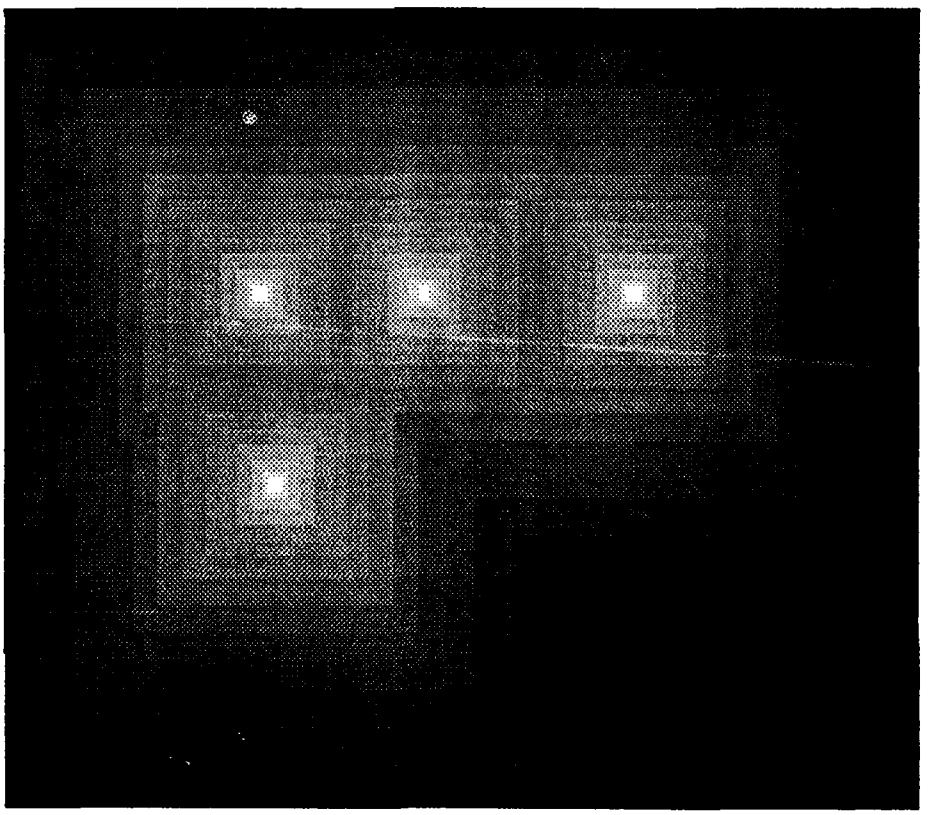
a. Acuity profile of integrated perception region after first registration.



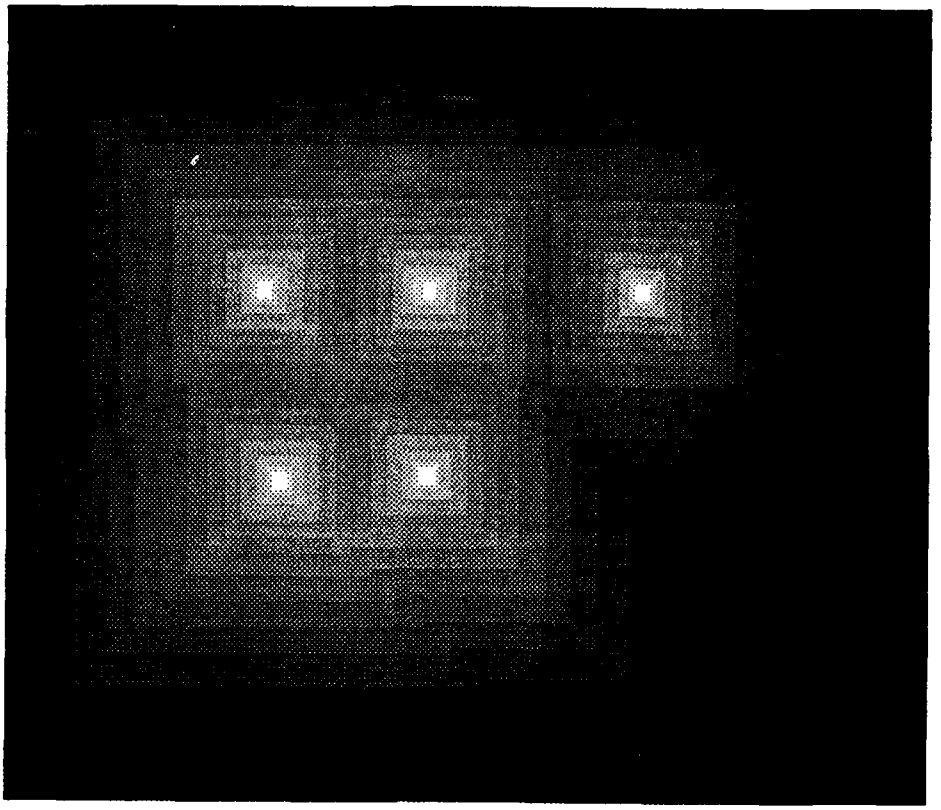
b. Acuity profile of integrated perception region after second registration.



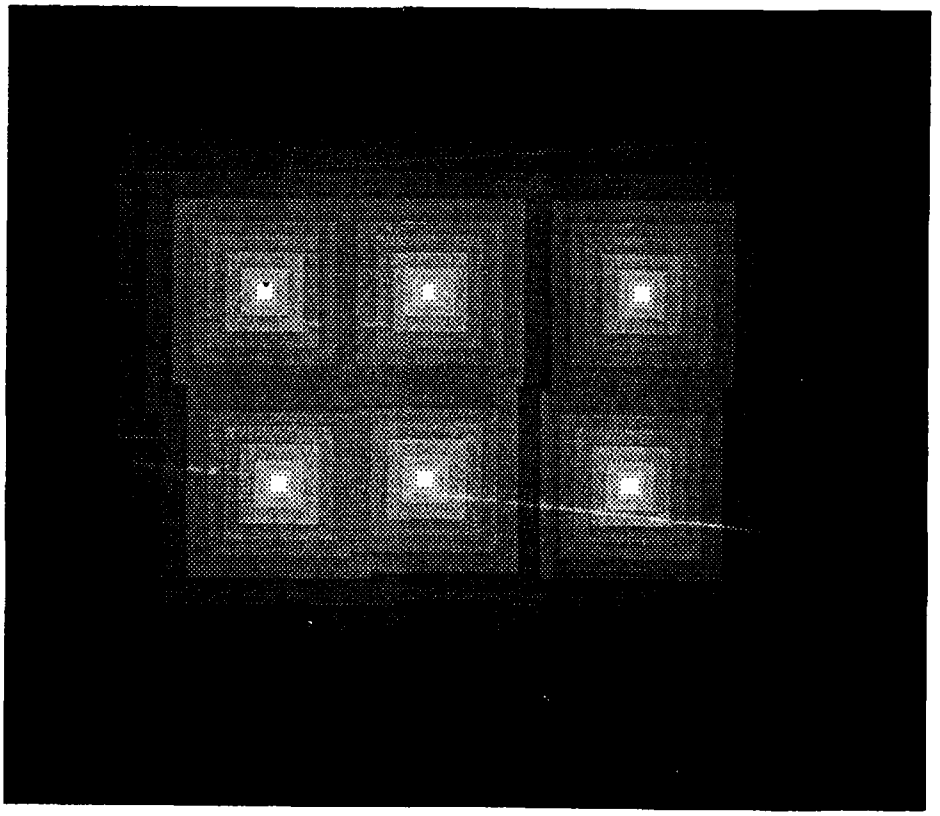
c. Acuity profile of integrated perception region after third registration.



d. Acuity profile of integrated perception region after fourth registration.



e. Acuity profile of integrated perception region after fifth registration.



f. Acuity profile of integrated perception region after sixth registration.

4.7 Data Fusion Accuracy

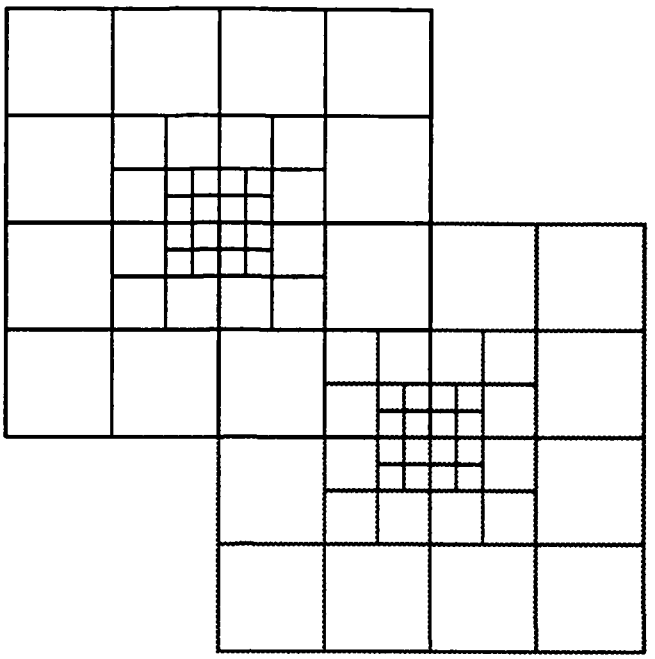
As proposed in Section 4.3, one way to quantify the data loss in the formation of a perception is by retrieving from the perception the best approximation to a previously integrated sensor frame, and measuring the difference between the approximation and the actual frame. The difference represents the information from the sensor frame that was not integrated into the perception.

In the absence of sensor noise, the value of a large rexel replaced by a group of smaller perfectly encompassed rexels is recoverable by taking the average of the smaller rexels. However, when the boundaries of the smaller rexels do not coincide with the boundaries of the large rexel being replaced, then the exact value of the large rexel cannot be recovered (Figure 4.7-1). This is because the small rexels extending beyond the boundary of the discarded large rexel average into their value scene luminosity not registered by the large rexel. Even though higher acuity data is obtained, edge information from the large rexel is lost. Nevertheless, this resegmentation is small when compared to the cumulative area of the small rexels entirely circumscribed by the large rexel boundary.

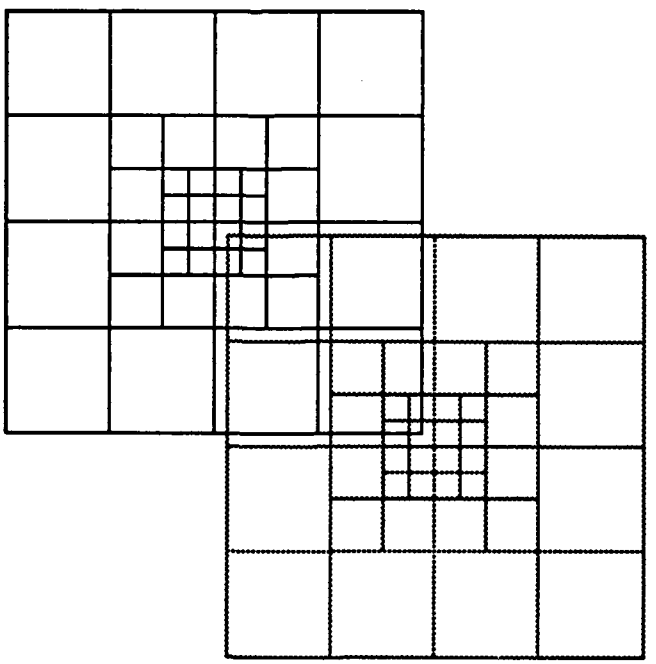
The value of a rexel is the average luminosity at a particular region of the scene. One approach to approximating a rexel value from the perception data is to average the luminosity perceived at the same region. Thus, the reconstruction of a frame of rexel data is obtained by foveating on the perception itself along the same optical axis. If the actual frame was previously integrated into the perception without data loss, and none of its rexels in the perception were altered by the integration of further sensor frames, then the approximation is exact.

A measure of lost frame data is the root-mean-square (RMS) difference between a particular frame registration and its reconstruction from the foveal sampling of the integrated perception at the same foveal axis. This reconstruction error varies with the initial perception state and the foveation sequence between the reference foveation and the time of reconstruction. For example, only the frame from the first registration can be reconstructed perfectly, and only before the second registration frame is integrated. This is because the first frame initializes the perception. Upon integrating the second frame, some of the data from the first frame is lost and may no longer be perfectly recoverable. Any frame other than the first one is not perfectly recoverable even immediately after integration

into the perception, because some of it is discarded in favor of higher acuity information already present in the perception database. Note that sensor noise does not factor into this test; the objective is to reconstruct sensor measurements (noise and all), not scene data.



a.



b.

Figure 4.7-1. Rexel alignment resulting in recoverable (a) and irrecoverable (b) fusion.

A question arises with the RMS reconstruction error measure: do we use the RMS difference between the rexel values or the pixel estimate values? Even though the format of the raw foveal sensor data is the rexel, its value alone does not represent all of its information; its size is of equal importance. The pixel estimates from a rexel are derived from the rexel's value and size, and represent the total rexel information. The RMS difference between the pixel estimate values is thus employed. The rexel value is thus effectively weighted by its size.

The second perception (integrated perception after the second registration) of the "Hello There" foveation sequence was foveally sampled at the location of the first registration. The RMS difference between the rexels of the first frame and its reconstruction is 0.032. Since the full 8 bit dynamic range is used (the difference between the black letters and the white background), the RMS error amounts to 0.0125 percent of the signal.

The reconstructed first registration frame is presented in Figure 4.7-2. Note its similarity with the actual first frame (Figure 4.5-2b). The absolute value difference between these two images is presented in Figure 4.7-3, where white represents zero difference (contrast enhanced for visibility). The difference appears only in the region of the perception where the second registration overwrites the first.

Likewise, the sixth perception of the foveation sequence was foveally sampled at the location of the first registration. In this case, the RMS difference between the rexels of the first frame and its reconstruction is 0.023, or 0.009%. In general, reconstruction error decrease as more frames are integrated (before or after the reference frame) because perception acuity improves, and foveally sampling the perception resembles more like foveally sampling the scene. The reconstructed first registration frame is presented in Figure 4.7-4. The difference between this image and the actual first frame is presented in Figure 4.7-5. Again, the difference appears only in the region of the perception where the the first registration was overwritten by the first foveal measurements that followed.

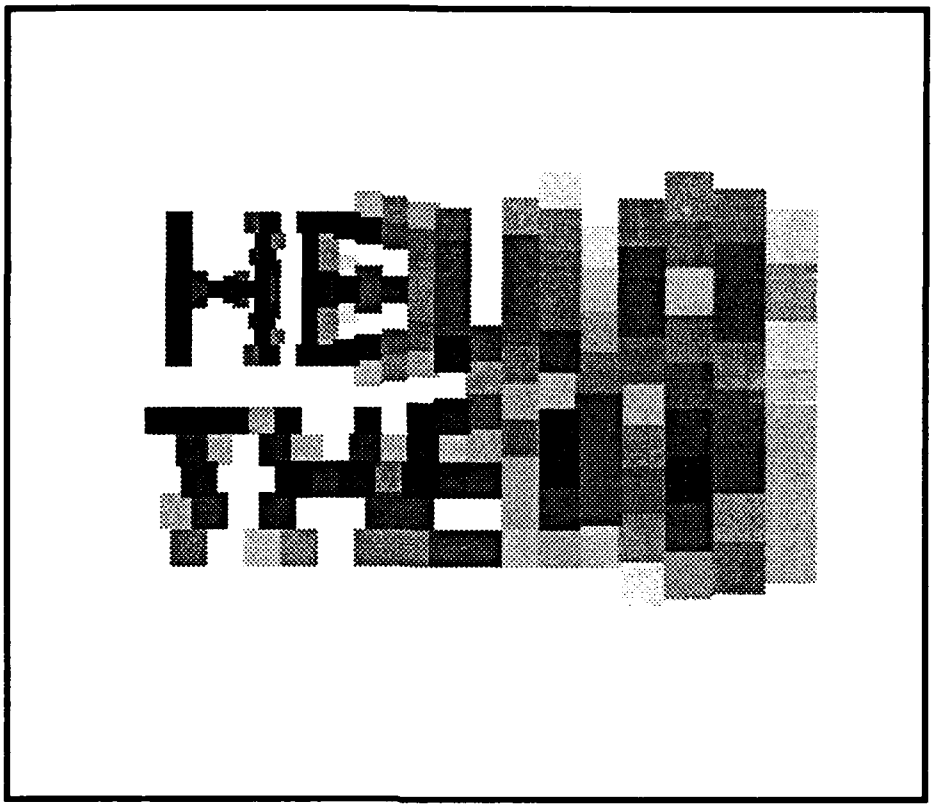


Figure 4.7-2. First sensor frame reconstructed from two registration perception.

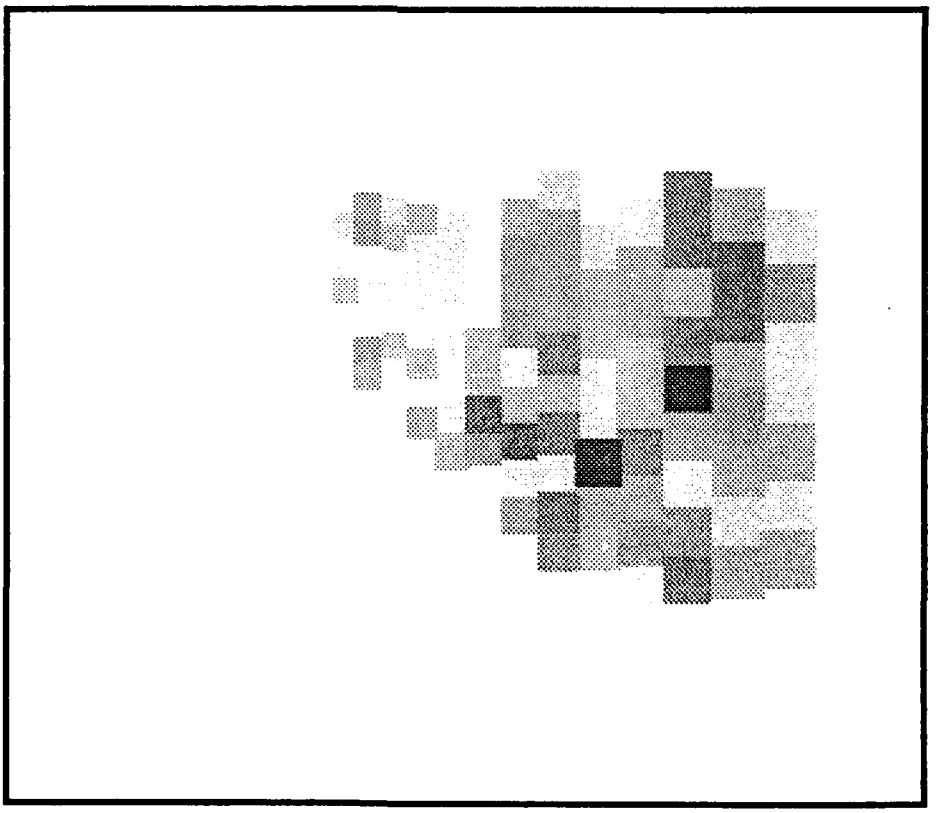


Figure 4.7-3. Error in reconstruction of first frame from second integrated perception.

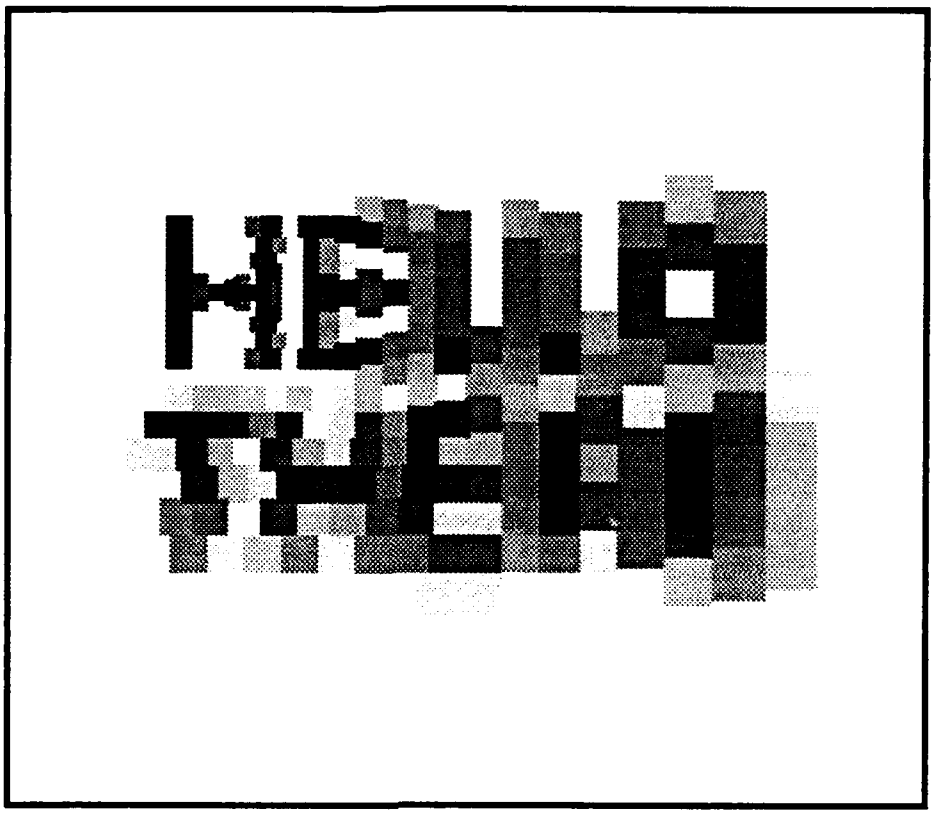


Figure 4.7-4. First sensor frame reconstructed from six registration perception.

Figure 4.7-5. Error in reconstruction of first frame from sixth integrated perception.

The text scene of the foveation example in Section 4.5 contains many sharp edges which are the first casualties of the assumption of pixel level statistics independence, as discussed earlier. A more natural scene is now employed. Consider a scene with little Encarnita (Figure 4.7-6a). The scene is foveally sampled at the locations indicated in Figure 4.7-6b, with the objective of resolving Encarnita's face and some background objects. The scene image contains 790×810=639,900 pixels.

The first frame and perception (Figure 4.7-c, d) register Encarnita with only 1603 rexels, yet with the necessary acuity to identify a smiling child. The two bathers in the background are better resolved in the next two foveations (Figures 4.7-e through h). Their features are small with respect to those of Encarnita but still occupy a region considerably larger than the fovea. Consequently, they are not as resolved with a single foveation as the face of Encarnita. Depending on the specific task, a few further interrogations of the bathers could be performed. The fourth foveation (Figures 4.7-i through j) resolves the windows of a hotel in the background. The features are even smaller than those of the bathers, but since the windows occupy a smaller region in the scene, they are registered more precisely and can even be counted after just the single interrogation. The last three foveations (Figures 4.7-k through p) better resolve Encarnita. The perception does not change significantly because there is little detail for the additional acuity to measure.

The evolution of the integrated perception is summarized in Table 4.7-1. The general decrease in the rate of perception growth is illustrated. As with the "Hello There" sequence, the first sensor frame of the "Encarnita" sequence was reconstructed by foveally sampling the second and last (seventh) integrated perception. The RMS difference between the first frame and its reconstruction from the second perception is 0.00321, or 0.00125% of signal value. The RMS difference between the first frame and its reconstruction from the seventh perception is 0.00456, or 0.00178%.

It is quantitatively seen that the integrated perception generated by the discard method retains the majority of measurement data. Sensor frames are accurately recoverable. These reconstruction error values are smaller than those of the "Hello There" sequence because the "Encarnita" scene has less high frequency edges which are lost through miscorrelation of rexels of different frames. The reconstruction from the second "Encarnita" perception is better than that from the seventh because more of the high acuity data from the first frame is retained by the former. The second perception differs from the first only in the far periphery when one of the bathers is resolved. The seventh perception,

on the other hand, has information from the last three foveations which are performed near the location of the first registration, and thus some of the higher acuity data from the first frame is replaced. The absolute value difference between the first frame and its reconstruction from the second and seventh perception is illustrated in Figures 4.7-7 and 4.7-8 respectively (contrast enhanced for visibility).

Registration	Rexels in sensor frame (in FOR)	Rexels dropped from perception	Frame rexels added to perception	Overall growth in perception	Size of perception in rexels
1	1603	0	1603	1603	1603
2	1468	327	919	592	2195
3	1361	260	570	310	2505
4	1571	362	761	399	2904
5	1521	396	679	283	3187
6	1644	232	447	215	3402
7	1602	284	486	202	3604

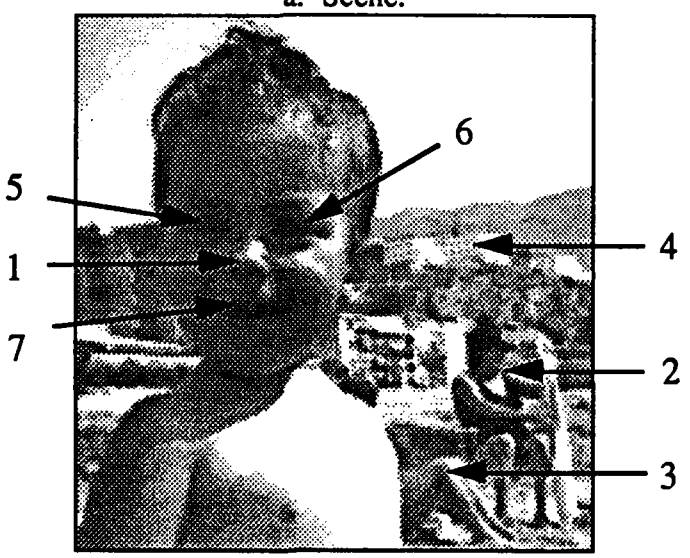
Table 4.7-1. Data retention and discard in "Encarnita" example of integrated perception evolvement.

4.8 Perception Database Growth

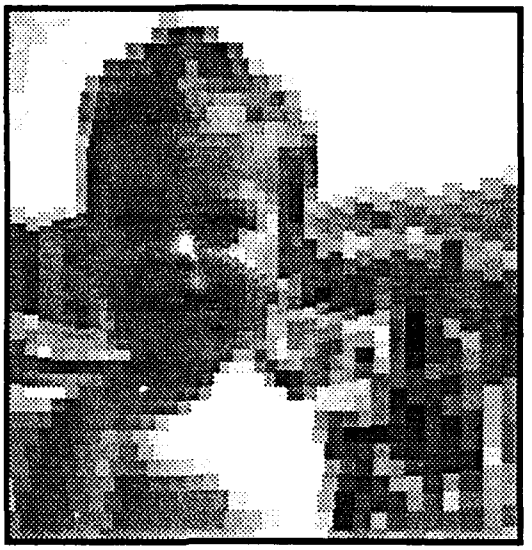
The process of generating a perception replaces low bandwidth data with high bandwidth data. Consequently, the perception grows with every integration of a sensor frame. Unlike the reversible data accumulation approach which generates a (non-integrated) perception that grows linearly with number of foveations, the rate of integrated perception growth generated by the discard method decreases with number of foveations. As more foveations are processed and the resolution of the overall perception increases, smaller portions of the sensor frames offer higher resolution data and are retained.

Figure 4.7-6. Encarnita perception sequence. (see image sequence below)

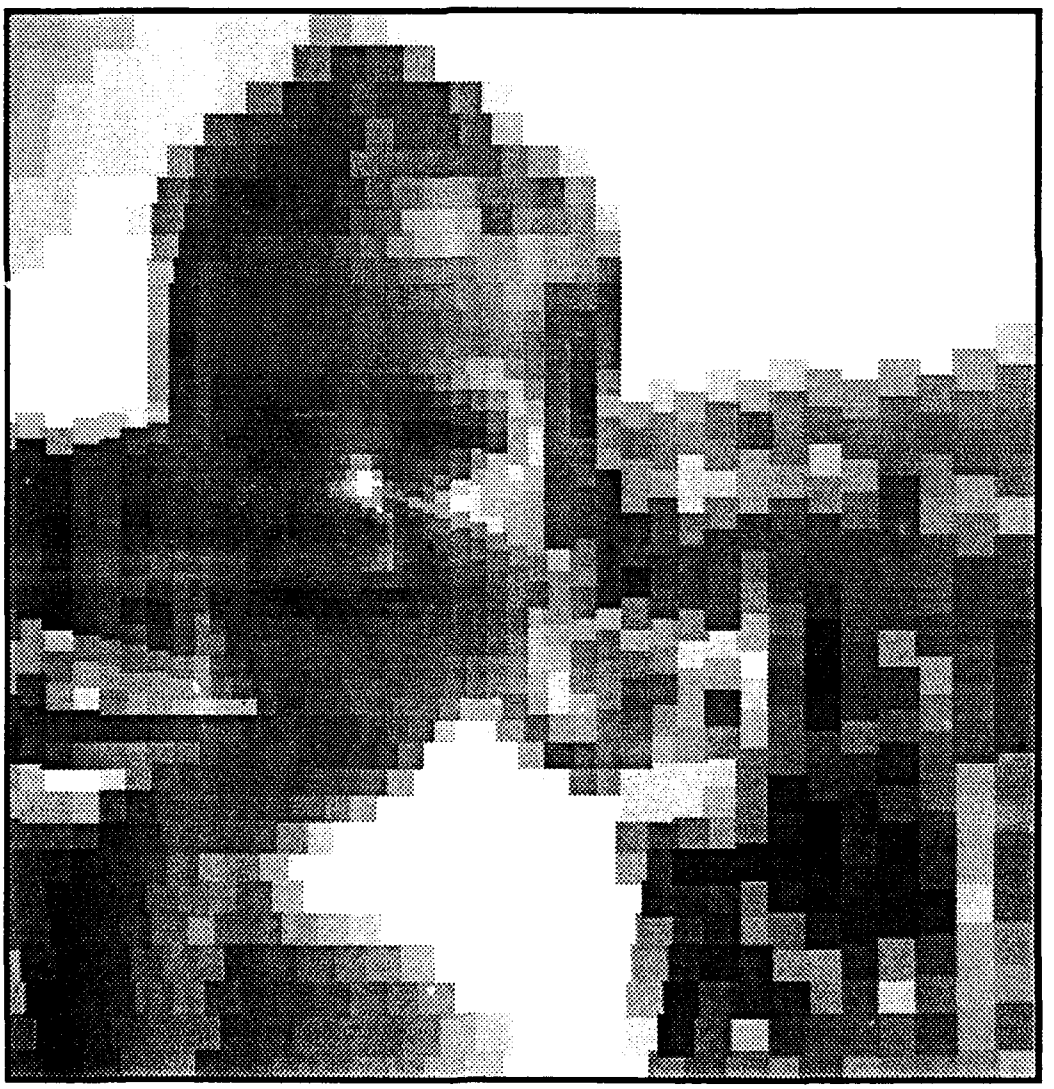
a. Scene.



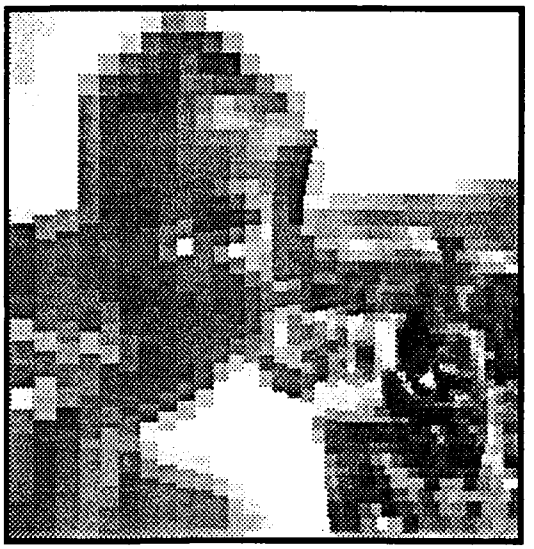
b. Foveation locations.



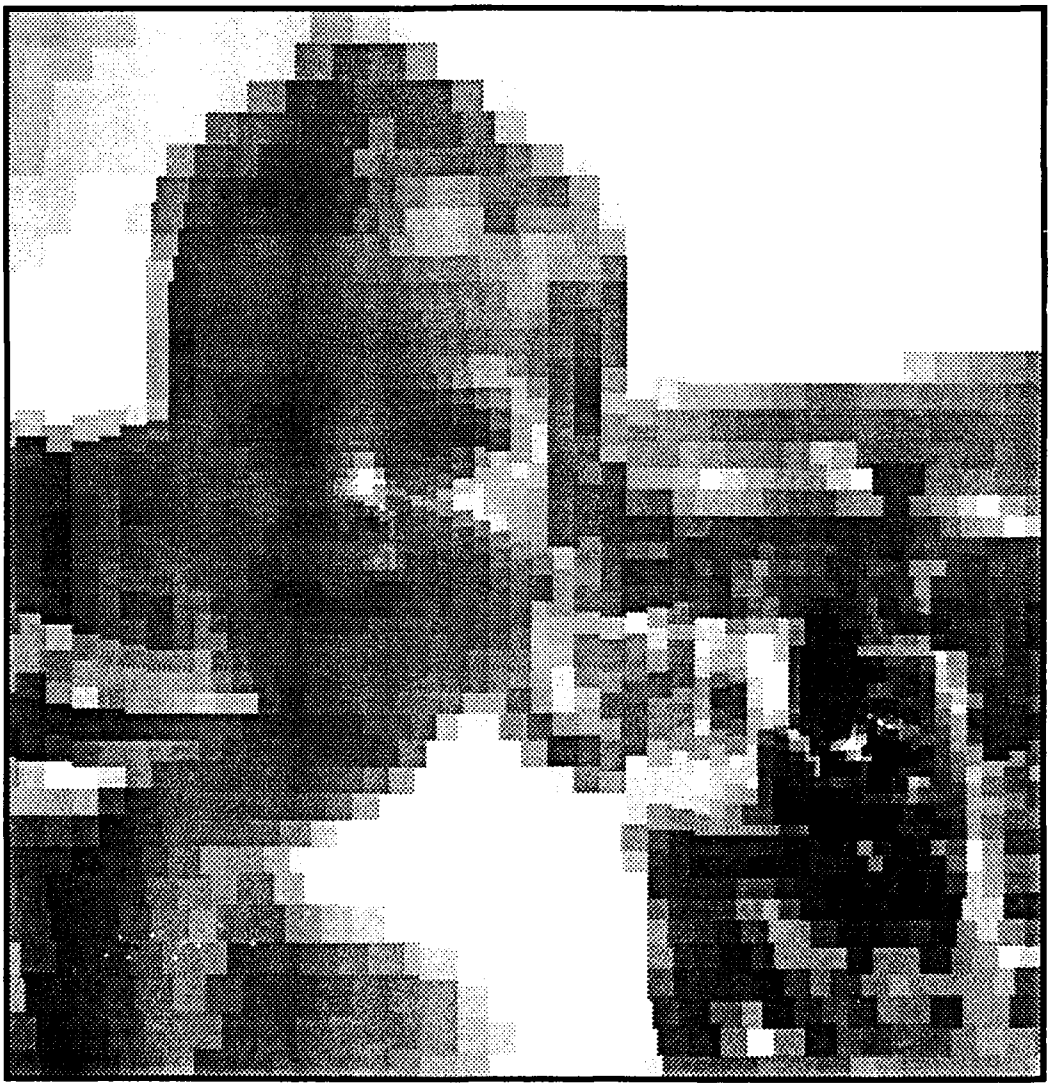
c. First sensor frame.



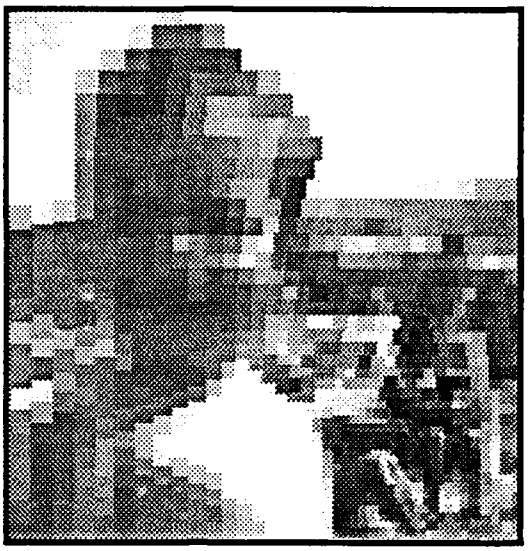
d. First integrated perception.



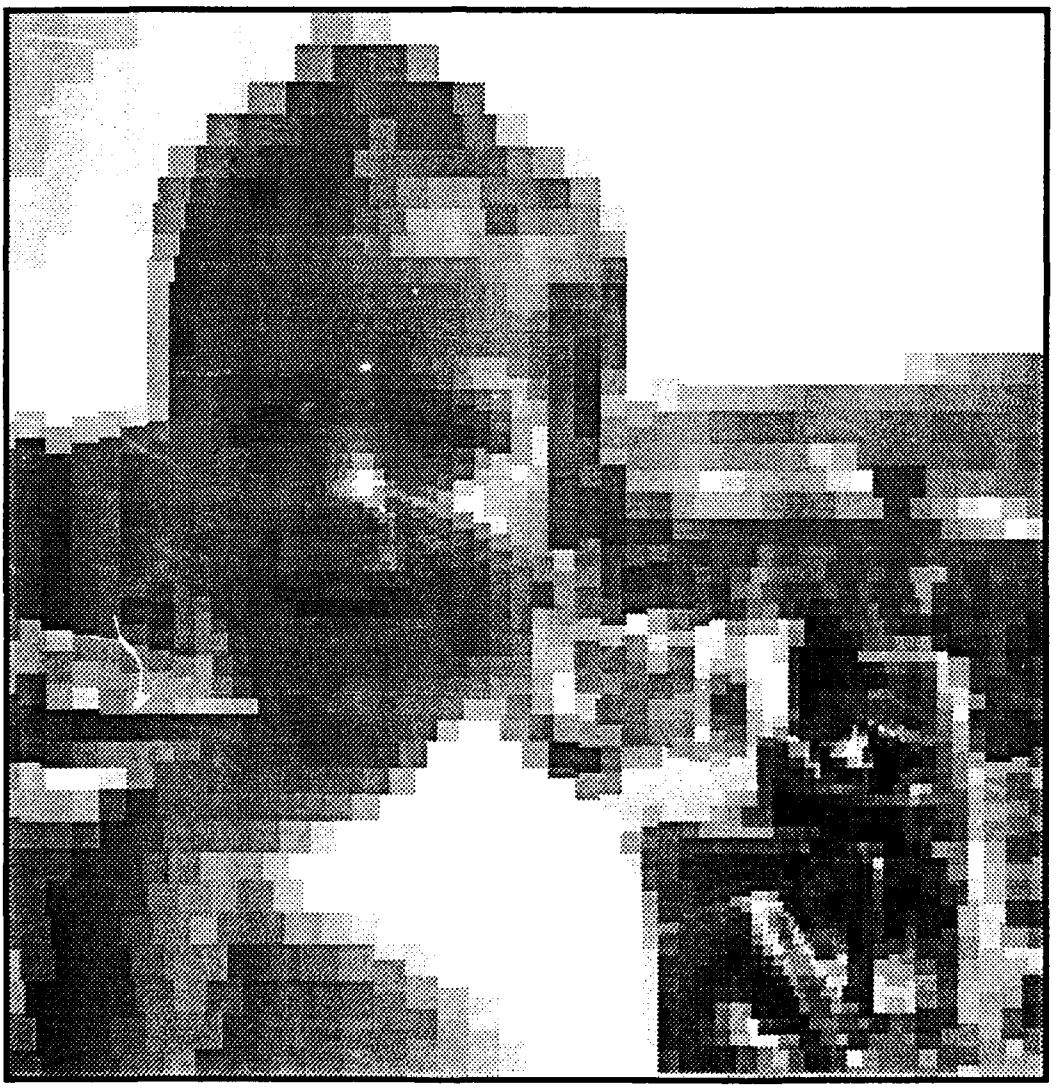
e. Second sensor frame.



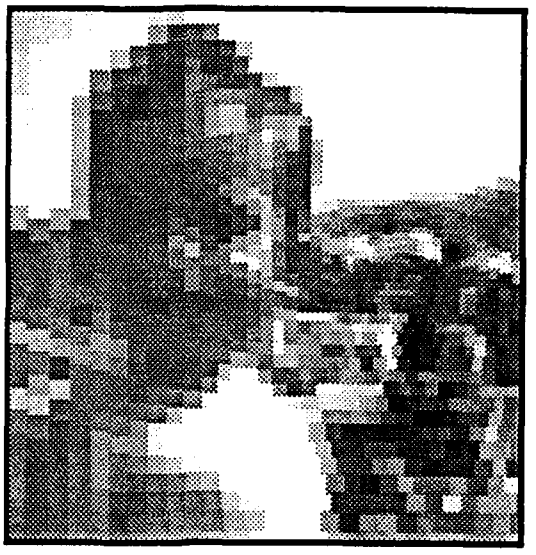
f. Second integrated perception.



g. Third sensor frame.

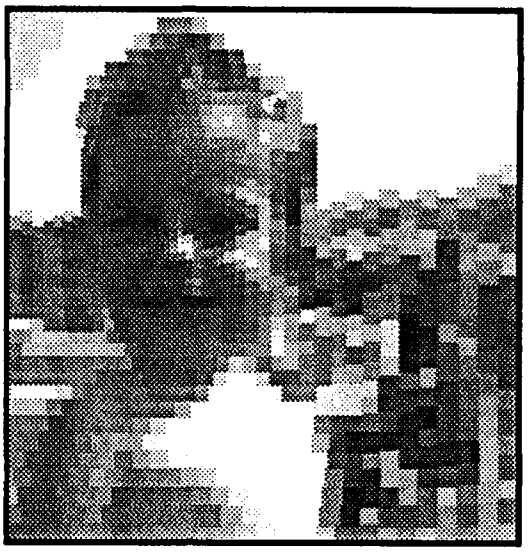


h. Third integrated perception.



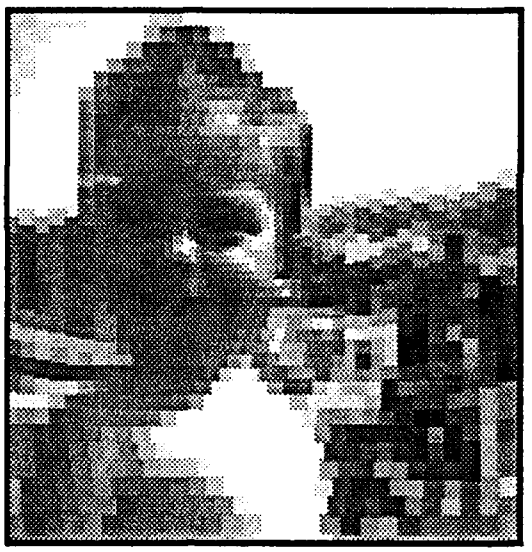
i. Fourth sensor frame.

j. Fourth integrated perception.



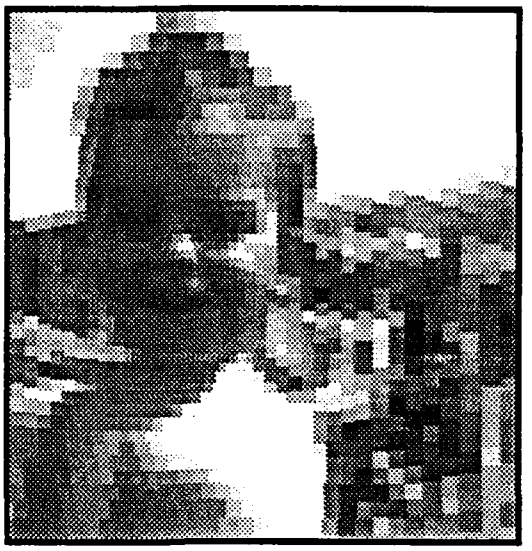
k. Fifth sensor frame.

l. Fifth integrated perception.



m. Sixth sensor frame.

n. Sixth integrated perception.



o. Seventh sensor frame.

p. Seventh integrated perception.

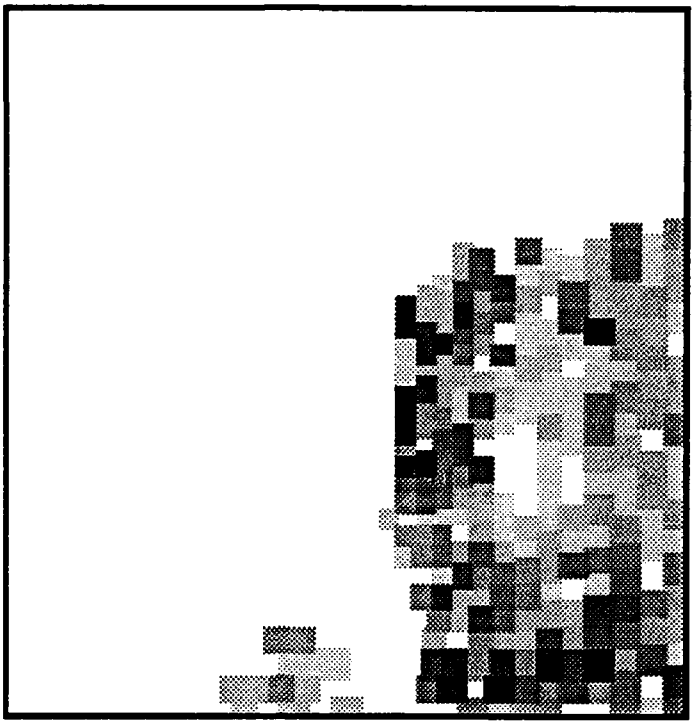


Figure 4.7-7. Error in reconstruction of first frame from first integrated perception.

Figure 4.7-8. Error in reconstruction of first frame from seventh integrated perception.

The rexel data from the first sensor frame is always adopted completely by the perception (except for rexels outside the field-of-regard). Beyond this initialization of the perception, the exact rate of increase in perception data is highly dependent on the location of the optical axis of each foveation, which in turn is highly dependent on the task being performed. Each foveation contributes a singly connected convex region of rexel data to the perception.[16] As previously illustrated (Figures 4.5-3 and 4.5-4), these regions can be frame labeled. Since only the highest acuity data from a foveation appears in the labeled region, the region boundaries can be considered as the decision boundaries of a minimum-distance two dimensional space linear classifier where the discriminant functions are shifted copies of the acuity function [Duda73]. In the typical case, where the acuity function is unimodal and the local maxima appears at the optical axis, the mean values of the discriminant functions are the optical axis themselves (Figure 4.5-1).

4.8.1 Perception Database Growth During Interrogation

Foveation locations are spaced closely together over time when the vision system is performing small feature analysis. The resulting frame labeled regions resemble radial slices emanating from optical axis locations. Radial slices of the perception are replaced by radial slices of high acuity sensor frame data. Since the vertex of a discarded perception slice is close to a previous optical axis, it has an acuity similar to but slightly less than the sensor slice, and the perception grows very little beyond initialization.

4.8.2 Perception Database Growth During Search

The tasks of object search and the interrogation of a large region requires the optical axes to be spaced apart. The maximum perception growth occurs when the optical axes are equidistant and the frame labeled regions are of the same size. In this case, after integrating

[16] The regions formed by the acuity criteria are technically not convex due to the ribbed edge effects from acuity discontinuity. The regions are convex if proximity to an optical axis is used as the data retention criteria. In noise free scenes without object motion, these two approaches produce the same perception.

the high resolution data from n sensor frames, the perception contains an amount of data equivalent to that of n frames with a field-of-view ε_n given by

$$n(\varepsilon_n)^2 = \varepsilon_0^2 = A_p \tag{4-39}$$

$$\varepsilon_n = \frac{\varepsilon_0}{\sqrt{n}} = \sqrt{\frac{A_p}{n}} \tag{4-40}$$

where ε_0 and A_p, are the linear dimension and area (in pixels) respectively of the total field-of-regard. The rexel count of a linear pattern $A_{r_l,n}$ with a field-of-view ε_n is obtained using (3-4):

$$A_{r_l,n} = 2\varepsilon_n = 2\sqrt{\frac{A_p}{n}} = \frac{A_{r_l}}{\sqrt{n}} \tag{4-41}$$

where A_{r_l} is the rexel count of a linear pattern with a field-of-view ε_0. The rexel count of an exponential $A_{r_e,n}$ with a field-of-view ε_n is obtained using (3-15):

$$A_{r_e,n} = 6\log_2(\varepsilon_n)^2 = 6\log_2\frac{A_p}{n} = A_{r_e} - 6\log_2 n \tag{4-42}$$

where A_{r_e} is the rexel count of an exponential pattern with a field-of-view ε_0. The worst case data size of a linear pattern perception after the integration of n registrations is expressed by

$$A_{percept,l} = nA_{r_l,n} = 2\sqrt{nA_p} = A_{r_l}\sqrt{n} \tag{4-43}$$

and its rate of growth is

$$\frac{d}{dn}A_{percept,l} = \sqrt{\frac{A_p}{n}} \tag{4-44}$$

The worst case perception size for the exponential pattern is

$$A_{percept,e} = nA_{r_e,n} = 6n\log_2\frac{A_p}{n} = n\left[A_{r_e} - 6\log_2 n\right] \tag{4-45}$$

with a rate of growth

$$\frac{d}{dn}A_{percept,e} = \frac{6}{\ln 2}\left[\ln\left(\frac{A_p}{n}\right) - 1\right] \tag{4-46}$$

The exponential foveal pattern, in addition to providing the smaller frame sizes, also features the smaller perception growth rate.

4.9 Pixel Versus Rexel Perception Format

Foveal sensor data can be processed in the rexel and/or pixel format. The rexel format consists of processing one datum per rexel. The implicit information accompanying the rexel luminosity value is the rexel location and its size (or resolution). The pixel format consists of the estimates of the values for the pixels covered by the rexel or region. The implicit information accompanying the pixel luminosity value is only the pixel location.

The advantage of pixel formatted data is that it can be stored in a conventional two dimensional uniform data structure, such as an array. This data structure explicitly stores only pixel values, but the ordering of pixels in the array maintains pixel locations implicitly. A disadvantage of pixel formatted rexel data is the overhead of redundant information, since the estimates of pixel values under each unisource region are identical.

The advantage of rexel formatted data is its small size. However, the storage of rexel data in a uniform two dimensional array is difficult because unisource regions can take on a wide variety of shapes. Rexel data can be stored in thinned uniform arrays, but this introduces database processing overhead. Hierarchical structures, such as a pyramid data structure, offer a more efficient approach to foveal sensor frame and perception storage. This architecture has three dimensions; two dimensions represent space as conventional two dimensional uniform arrays, and a third dimension represents resolution. Just as uniform arrays implicitly represent the two dimensional position of pixel information, pyramid data structures implicitly represent the three dimensions of implicit information accompanying each rexel value.

Chapter 5

Gaze Control Strategies

5.1 Introduction to the Foveal Gaze Control

Gaze control strategies are very task dependent, since different tasks prioritize information differently. Nevertheless, the concept of maximizing relevant information is common to all strategies. The low acuity peripheral perception of scene features provides limited information on the features themselves, but it does present cues on the global state of the environment in a compact fashion. These cues are exploited by the foveal control strategy to hypothesize the state of nature (e.g., scene grey levels, or the classification and location of objects in the scene). From this hypothesis, gaze angles for saccades are selected which maximize the expected *relevant* information in the sensor data.

The process of foveation, which includes the selection of the new gaze angle, is addressed here as one of optimal statistical control. In this approach, the integrated perception is labeled as the state of the system. A scalar figure of merit representing the content of relevant information in the perception is maximized in the presence of system perturbations (e.g., sensor noise, gimbal dynamics, etc.). A feature of the foveal machine vision system which distinguishes it from conventional control systems is the controllable (via gaze angle) reduction in the order of the input vector (scene luminoscity) due to rexel averaging.

The top level view of the foveal machine vision system as a closed loop system is given in Figure 5.1-1. The space variant sampling and the integrated perception have been discussed in previous chapters. The data coding block derives feature information from the raw sensor data such as scene pixel estimates driving a low level perception. The foveation controller implements the gaze control strategy. The driver of the control algorithm is a task dependent cost function, which attempts to maximize the information content in the next sensor frame. The cost function is used by the foveation controller, which generates a hypothesis of the scene based on perceived data (*a priori* information), and selects the new

gaze angle. The next sensor frame provides *a posteriori* information which updates the perception and refines the hypothesis for the selection of the next gaze angle.

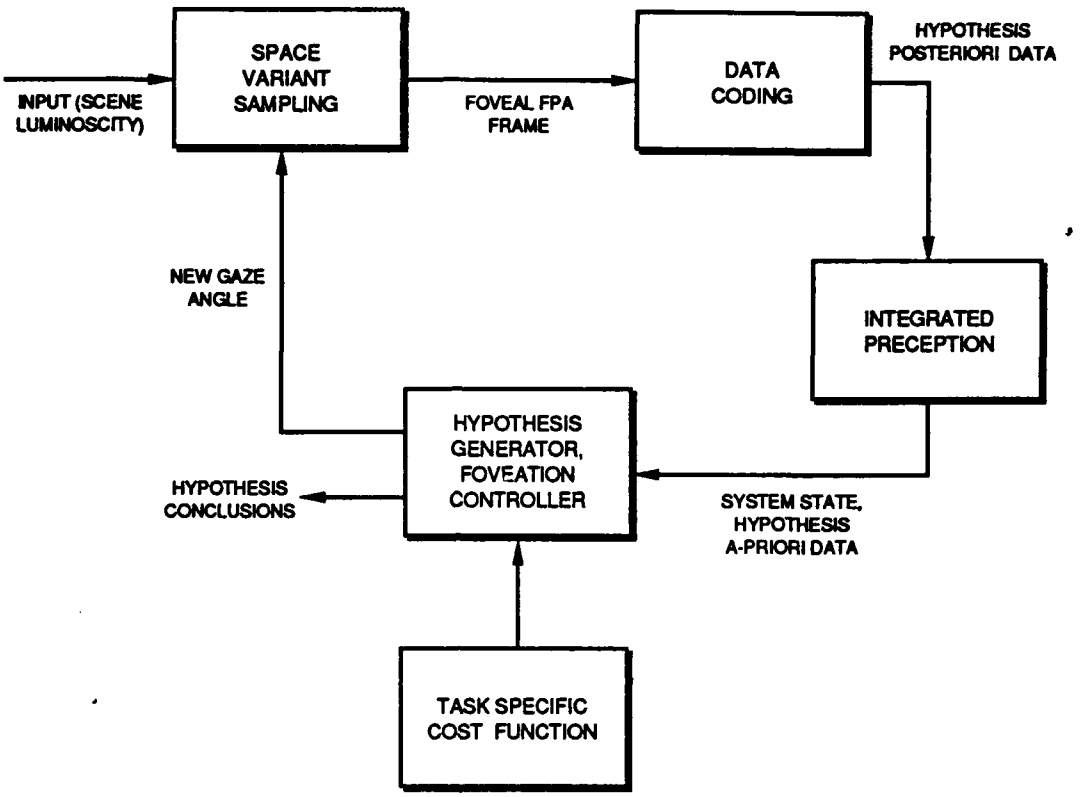


Figure 5.1-1. Control system model of foveal machine vision system.

Two general gaze control strategies are presented in this chapter. Each defines a particular mode of operation in the foveal system. The first, called the *survey mode*, is a global learning strategy. This strategy is employed to obtain relevant information on the overall context of the scene, and can serve as an initial mode of operation for many vision tasks. One implementation of the survey mode strategy is the minimization of the entropy of the system hypotheses. Hypothesis entropy is appropriate for this strategy because it measures the total amount of information perceived by the foveal system relevant to its hypotheses, and thus to the task of global context learning. When in the survey mode, a foveal system selects the sensor gaze angle, or sequence of gaze angles, expected to produce the sensor frame(s) which minimize this entropy (i.e., offer the highest content of relevant information).

The second gaze control strategy is called the *interrogation mode*. As opposed to the global learning operation of the survey mode, the interrogation mode allocates resolution resources to a specific cue or localized cue cluster. This strategy is employed when the foveal system is analyzing a particular scene feature, or equivalently, testing a particular hypothesis. One implementation of the interrogation mode strategy is the maximization of the hypothesis likelihood ratio. The likelihood ratio is appropriate for this strategy because it measures the total amount of information perceived by the foveal system relevant to the particular hypotheses under test, and thus to the task of feature analysis. When in the interrogation mode, the system foveates directly to the cue or cur cluster centroid.

The foveal system can dynamically switch from one gaze control strategy to another as the immediate objective of the system changes (within the overall goal of completing the vision task). For example, the initial mode of the foveal system can be the survey mode, in which the system learns about the scene in general and detects cues. When a particularly interesting cue is detected, the system may switch to the interrogation mode to further resolve and analyze the cue. Upon completion of the cue analysis, the system would revert back to the survey mode and search for new cues, particularly in regions in the field-of-regard where there has not yet been any foveation and which are represented in the integrated perception with low acuity.

5.2 Expected Entropy Minimization

A simple cost function is the information content of the perception. The objective of a foveation controller in the survey mode is to maximize this value. A popular measure of information content is entropy [Hammi90], which in terms of perception data is given by

$$E = -\sum_{\text{all } \Re_i \in \Re} P(\Re_i|\Im)\log_2\left(P(\Re_i|\Im)\right) \tag{5-1}$$

where $\Im$ is the set of all data in the perception database, $\Re$ is a set of exhaustive and mutually exclusive hypotheses $\Re=\{\Re_1, \Re_2, ..., \Re_{max}\}$ on specific features or events,

$$\sum_{\text{all } \Re_i \in \Re} P_{\Re}(\Re_i) = 1 \tag{5-2}$$

and $P(\Re|\Im)$ is the *a posteriori* probability of this hypothesis given all the information perceived so far. In this expression, entropy is measured in bits. The features or events employed depend on the task being performed. For example, in target localization, $\Re_i$ can be the presence of a target at the i'th pixel of the field-of-regard.

Entropy is a measure of the collective ambiguity of all the hypotheses: when E=0, the hypotheses are resolved (confirmed or denied) deterministically. In target localization, zero entropy means the system is certain as to the location of targets in the field-of-regard. With the introduction of clutter and sensor noise, entropy will never be zero (there will always be some chance of a false detection and false negative), but thresholds can be arbitrarily set to obtain desired confidence values.

The objective of the foveation controller implementing the expected entropy minimization gaze control strategy is to select a gaze angle which produces new observations which in turn, when integrated into the perception, minimize the entropy of the task hypotheses. Thus, the gaze angle minimizing the expected *a posteriori* entropy is selected. Since perception ambiguity is reduced by sensor information, minimizing perception entropy is equivalent to maximizing the information content in the sensor frame.

The $P(\Re_i|\Im)$ term is obtained from *a priori* knowledge on the event and the information received so far using Bayes rule

$$P(\Re_i|\Im) = \frac{P(\Re_i)P(\Im|\Re_i)}{P(\Im)} \tag{5-3}$$

where $P(\Re_i)$ is the *a priori* probability for hypothesized event $\Re_i$, $P(\Im|\Re_i)$ is the probability of the perceived information assuming the hypothesis is true, and $P(\Im)$ is the probability of the perceived information. This leads to one significant drawback of using entropy as a cost function, namely that *a priori* information which may not be reliably available is required.

Another problem with Bayes rule is its awkward performance when there are many hypotheses and the value of any particular $P(\Re_i)$ is very low. Such is the case when localizing a small number of unresolved targets in a large field-of-regard. In this case, the ratio of likelihoods

$$\frac{P(\Im|\Re_i)}{P(\Im)} = \frac{P(\Im|\Re_i)}{\sum_i P(\Im|\Re_i)P(\Re_i)} \tag{5-4}$$

must be very great under observations which corroborate a hypothesis in order to scale the low *a priori* probability value $P(\Re_i)$ to an *a posteriori* probability of significant value. This can be difficult when the applicable data from the perception was measured with low acuity because the data is ambiguous, or equivalently, $P(\Im|\Re_i)$ features a large variance due to the averaging by large rexels of significant (and possibly overwhelming) noise and clutter in $\Im$. This agrees with intuition; confirmation of hypotheses on detailed features should not be expected with low acuity data obtained with peripheral vision.

5.2.1 Optimum Entropy Minimization

An algorithm for optimum entropy minimization is presented in this section. This approach employs the reversible state of nature integrated perception presented in Section 4.3. Due to the enormous state space employed, this algorithm is not feasible for applications of significant size (field-of-regard, signal quantization resolution). The algorithm is nevertheless presented as a reference to functional approximations discussed in following sections which reduce the state space and algorithm complexity to feasible levels.

The process of computing the expected hypothesis entropy after foveating to some location λ requires predicting the *a posteriori* state of the integrated perception. This prediction, in turn, requires a prediction of the data such a foveation will produce. To accomplish all of this, an exhaustive set of cases must be formed, each postulating that a specific state of nature is true. The results can then be combined into a statistical moment.

As in Section 4.3, let x_i be a state of nature in I^{N^2} space, where I indicates the integer domain of zero to some maximum number I_{max} -1 which represents the quantization of measurements, and N^2 is the number of pixels in the field-of-regard. Thus, there are $(I_{max})^{N^2}$ unique states of nature each which can be represented by the array of N^2 pixel values of the corresponding field-of-regard. The reversible state of nature integrated perception assigns a probability to each x_i, or equivalently, forms the distribution function

$$P_x(\omega) = P(\text{scene is } x_\omega) \qquad \omega = 1 \ldots (I_{\max})^{N^2} \tag{5-5}$$

The vales of the hypotheses can be binary functions (i.e., taking on the value true or false) of the state of nature

$$\Re_i = f_{\Re_i}(\omega) \tag{5-6}$$

The expected value of a hypothesis after foveating to location λ is a function of the expected perception. The computation of the expected perception requires computing the expected rexel data from the foveation to λ.

The rexel data $\overline{R}_{\lambda|x_i}$ expected from a foveation to λ given that the state of nature is x_i is obtained by simulating the foveal sampling process; the scene is postulated, and the gaze angle and foveal sensor geometry are known. The probability distribution of the rexel data conditioned on the scene identity x_ω is given by (4-10), (4-11), and

$$p\left(\overline{R}_{\lambda|x_i}\middle|\omega\right) = p\left(r_{1,\lambda}, r_{2,\lambda}, \ldots, r_{m_\lambda,\lambda}\middle|\omega\right) = p\left(r_{1,\lambda}\middle|\omega\right)p\left(r_{2,\lambda}\middle|\omega\right)\cdots p\left(r_{m_\lambda,\lambda}\middle|\omega\right) \tag{5-7}$$

where $r_{i,\lambda}$ is the i'th rexel of the frame with the optical axis at location λ of postulated state x_i, and m_λ is the number of rexels in the data frame within the field-of-regard.

Bayes rule provides the probability distribution on the states of nature conditioned on the rexel data, giving

$$P_x\left(\omega|\overline{R}_{\lambda|x_i}\right) = \frac{p\left(\overline{R}_{\lambda|x_i}\middle|\omega\right)}{p\left(\overline{R}_{\lambda|x_i}\right)} P_x(\omega) \tag{5-8}$$

where $p\left(\overline{R}_{\lambda|x_i}\right)$ is the probability of the expected data from the postulated scene based on some *a priori* model of the scene and sensor. With measurements from n previous registrations, (5-8) becomes

$$P_x\left(\omega|R_1, \ldots, R_n, \overline{R}_{\lambda|x_i}\right) = P_x(\omega)\frac{p\left(\overline{R}_{\lambda|x_i}\middle|\omega\right)}{p\left(\overline{R}_{\lambda|x_i}|R_1, \ldots, R_n\right)}\prod_{t=1}^{n}\frac{p\left(R_t|\omega\right)}{p\left(R_t|R_1, \ldots, R_{t-1}\right)} \tag{5-9}$$

where R_k is the set of rexel data from the k'th registration, and the product term represents the Bayesian learning updates from the registrations. An *a posteriori* hypothesis probability conditional on the next measurement data (i.e., conditional on foveating to λ in postulated scene x_i) is

$$P\left(\Re_i \middle| R_1,\ldots,R_n,\overline{R}_{\lambda|x_i}\right) = P\left(f_{\Re_i}\left(\omega \middle| R_1,\ldots,R_n,\overline{R}_{\lambda|x_i}\right)\right) \tag{5-10}$$

and the expected conditional hypothesis entropy becomes

$$E_{\lambda|x_i} = -\sum_{\text{all } \Re_i \in \Re} P\left(f_{\Re_i}\left(\omega \middle| R_1,\ldots,R_n,\overline{R}_{\lambda|x_i}\right)\right)\log_2\left[P\left(f_{\Re_i}\left(\omega \middle| R_1,\ldots,R_n,\overline{R}_{\lambda|x_i}\right)\right)\right] \tag{5-11}$$

The expected entropy value without the assumption on the state of nature is the sum of the conditional entropies weighted by the *a posteriori* probability on the state itself

$$E_\lambda = \sum_{i=1}^{(I_{max})^{N^2}} P_x\left(x_i \middle| R_1,\ldots,R_n\right)E_{\lambda|x_i} \tag{5-12}$$

The selected location λ^* is that which minimizes E_λ:

$$\lambda^* = \left\{\lambda^* \middle| E_{\lambda^*} = \min_\lambda E_\lambda\right\} \tag{5-13}$$

The minimization process can be constrained by further considerations, such as limitations of the gimbal mechanism (e.g., angular velocity, acceleration, and wear attributes). In this case, the selection of λ^* is a function of the gaze history of the system Λ and the gimbal dynamics G in addition to perception (hypothesis) entropy:

$$\lambda^* = \left\{\lambda^* \middle| f\left[E_{\lambda^*}, G\left(\lambda^*,\Lambda\right)\right] = \min_\lambda f\left[E_\lambda, G(\lambda,\Lambda)\right]\right\} \tag{5-14}$$

The above strategy solves for a single optimum optical axis location, which is then used for the next foveation. Control strategies such as this one, which solve for a sequence of control values (the sequence of λ^* over time) one value at a time by optimizing over one time step into the future, are called *myopic* strategies. The myopic strategy provides an optimal solution for the single time step, but as a component of a sequence of control values, the product of the myopic strategy can converge to a local minimum as opposed to a global minimum.

An alternative to the myopic strategy is the *n*-step look-ahead strategy, which solves for the optimum sequence of *n* control values. This strategy reduces to the myopic case when n=1. As in the myopic case, the selection of an optimum sequence of *n* foveations $\Lambda^* = \{\lambda_1^*, \lambda_2^*, \cdots, \lambda_n^*\}$ requires the estimation of the *a posteriori* probabilities of the states of nature after the sequence is performed. Since the ordering of foveations is irrelevant to the Bayesian learning process (4-13), the number of sequences s which must be considered is

$$s = \binom{N^2}{n} = \frac{N^2!}{n!\left(N^2 - n\right)!} \tag{5-15}$$

The rexel data expected from a sequence of foveations Λ assuming some state of nature x_i, expressed as

$$\overline{R}_{\Lambda|x_i} = \left\{\overline{R}_{\lambda_1|x_i}, \overline{R}_{\lambda_2|x_i}, \ldots, \overline{R}_{\lambda_n|x_i}\right\} \tag{5-16}$$

is obtained by simulating the foveal sampling process. From (5-7), the conditional statistics on the data given the scene identity is

$$p\left(\overline{R}_{\lambda_t|x_i}\middle|\omega\right) = p\left(r_{1,\lambda_t}, r_{2,\lambda_t}, \ldots, r_{m_{\lambda_t},\lambda_t}\middle|\omega\right) = p\left(r_{1,\lambda_t}\middle|\omega\right)p\left(r_{2,\lambda_t}\middle|\omega\right)\cdots p\left(r_{m_{\lambda_t},\lambda_t}\middle|\omega\right) \qquad t = 1,\ldots,n \tag{5-17}$$

and (4-10), (4-11), where m_{λ_t} is the number of rexels in the t'th frame within the field-of-regard. The expected *a posteriori* probability on the states of nature after performing Λ (postulating the scene to be x_i) is

$$\begin{aligned} P_x\left(\omega|\overline{R}_{\Lambda|x_i}\right) &= P_x\left(\omega|\overline{R}_{\lambda_1|x_i}, \overline{R}_{\lambda_2|x_i}, \ldots, \overline{R}_{\lambda_n|x_i}\right) \\ &= P_x(\omega)\frac{p\left(\overline{R}_{\Lambda|x_i}\middle|\omega\right)}{p\left(\overline{R}_{\Lambda|x_i}\right)} = P_x(\omega)\prod_{t=1}^{n}\frac{p\left(\overline{R}_{\lambda_t|x_i}\middle|\omega\right)}{p\left(\overline{R}_{\lambda_t|x_i}\middle|\overline{R}_{\lambda_1|x_i}, \ldots, \overline{R}_{\lambda_{t-1}|x_i}\right)} \end{aligned} \tag{5-18}$$

where the product term represents the iterative Bayesian learning from the sequence of foveations Λ (4-13). With measurements from *m* previous registrations, (5-8) becomes

$$P_x\left(\omega|R_1,\ldots,R_m,\overline{R}_{\Lambda|x_i}\right)=P_x\left(\omega|R_1,\ldots,R_m,\overline{R}_{\lambda_1|x_i},\overline{R}_{\lambda_2|x_i},\ldots,\overline{R}_{\lambda_n|x_i}\right)$$

$$=P_x(\omega)\left(\prod_{t=1}^{n}\frac{p\left(\overline{R}_{\lambda_t|x_i}\middle|\omega\right)}{p\left(\overline{R}_{\lambda_t|x_i}\middle|\overline{R}_{\lambda_1|x_i},\ldots,\overline{R}_{\lambda_{t-1}|x_i},R_1,\ldots,R_m\right)}\right)\left(\prod_{t=1}^{m}\frac{p\left(R_t|\omega\right)}{p\left(R_t|R_1,\ldots,R_{t-1}\right)}\right) \quad (5\text{-}19)$$

As with (5-10) and (5-11), the *a posteriori* conditional hypothesis probability is

$$P\left(\Re_i|R_1,\ldots,R_m,\overline{R}_{\Lambda|x_i}\right)=P\left(f_{\Re_i}\left(\omega|R_1,\ldots,R_m,\overline{R}_{\lambda_1|x_i},\overline{R}_{\lambda_2|x_i},\ldots,\overline{R}_{\lambda_n|x_i}\right)\right) \quad (5\text{-}20)$$

the conditional hypothesis entropy is

$$E_{\Lambda|x_i}=-\sum_{\text{all }\Re_i\in\Re}P\left(f_{\Re_i}\left(\omega|R_1,\ldots,R_n,\overline{R}_{\Lambda|x_i}\right)\right)\log_2\left[P\left(f_{\Re_i}\left(\omega|R_1,\ldots,R_n,\overline{R}_{\Lambda|x_i}\right)\right)\right] \quad (5\text{-}21)$$

and the expected entropy value without the assumption on the state of nature is

$$E_\Lambda=\sum_{i=1}^{(I_{max})^{N^2}}P_x\left(x_i|R_1,\ldots,R_n\right)E_{\Lambda|x_i} \quad (5\text{-}22)$$

The selected foveation sequence Λ^* is that which minimizes E_λ:

$$\Lambda^*=\left\{\Lambda^*\middle|E_{\Lambda^*}=\min_{\Lambda}E_\Lambda\right\} \quad (5\text{-}23)$$

It is quite obvious that the direct implementation of the ideal hypothesis entropy is unfeasible for fields-of-regard of any significant size. Table 5.2.1 summarizes the operations for n-step optimization and its computational requirements. Consider a modest 512×512 field-of-regard, eight bit signal quantization, and $m_{\lambda_{n,s}}<<N^2$, where $m_{\lambda_{n,s}}$ is the number of rexels generated in a foveation to the n'th location of the s'th sequence. The number of values processed by the second algorithm step when solving for a sequence of 10 foveations is $256^{2\times262144}\times4.2\times10^{47}\times2621440 \cong 10^{1,262,612}\times10^{47}\times10^{7}=10^{1,262,666}$ (current estimates place the number of subatomic particles in the universe at 10^{80}).

	Optimum n-step foveation sequence selection using minimum hypothesis entropy criterion.	Computational requirements
1.	Generate rexel data $\overline{R}_{\Lambda\|x_i}$ assuming state of nature x_i for all states i=1...$(I_{max})^{N^2}$ and for all sequences	$(I_{max})^{N^2}$ foveations simulated.
2.	Compute statistics on the conditional expected data $p(\overline{R}_{\lambda_t\|x_i})$ and $p(\overline{R}_{\lambda_t\|x_i}\|\omega)$ for all s sequences, t=1...n, and i,ω=1...$(I_{max})^{N^2}$	$(I_{max})^{2(N^2)}\binom{N^2}{n}\sum_{i=1}^{n}(m_{\lambda_{\omega t}}+N^2)$ values processed and evaluations of sensor noise distribution function
3.	Compute the *a posteriori* conditional probability on the states of nature $P_x(\omega\|R_1,\ldots,R_m,\overline{R}_{\Lambda\|x_i})$ for all s sequences and ω=1...$(I_{max})^{N^2}$	$n\binom{N^2}{n}(I_{max})^{N^2}$ individual Bayesian update terms computed
4.	Compute the conditional hypothesis entropy $E_{\Lambda\|x_i}$ after each of s sequences for all H hypotheses	processing of H self-entropy terms
5.	Compute the expected hypothesis entropy E_Λ after each of s sequences for all H hypotheses	weighted sum of $H(I_{max})^N$ terms
6.	Selection of sequence associated with minimum expected entropy	search for minimum of $(I_{max})^{N^2}$ values

Table 5.2.1-1. Optimum hypothesis entropy reducing algorithm for n-step foveation sequence selection.

5.2.2 Optimum Entropy Minimization in a Reduced Environment

In this section we observe the operation of the optimum selection of foveation locations (in the sense of minimum expected *a posteriori* hypothesis entropy) for the task of unresolved target localization. The foveal system will remain in the survey mode until the task is complete. The size and quantization of the scene are substantially reduced for reasons of experimental tractability. Consider a one dimensional scene containing one target and static clutter. The entire scene is defined to be the field-of-regard, and consists

of a linear array of N clutter pixels with equiprobable values of 0 and 1, and one stationary target of value greater than 1. Hence, excluding the target pixel, I_{max}=2.

The foveal sensor consists of three rexels of size 2×1 (r_1), 1×1 (r_2), and 2×1 (r_3) pixels. The foveal axis, located at the 1×1 rexel, can be aligned to any pixel in the scene. To reduce the search space for the algorithm, sensor noise variance is assumed negligible. To circumvent the handling of rexels or part thereof falling outside the field-of-regard, the scene is made circular such that space wraps around and the two ends of the linear array are connected (Figure 5.2.2-1). Thus, the component of a rexel extending beyond the last pixel of the scene covers the first pixel(s).

The value of the target, the probability distribution of the clutter, and the fact that there is only one target in the scene (field-of-regard), are assumed known by the system. No *a priori* information on target location is provided. There are then N hypotheses

H_1: the target is in scene pixel #1

H_2: the target is in scene pixel #2

$\vdots$

H_N: the target is in scene pixel #N

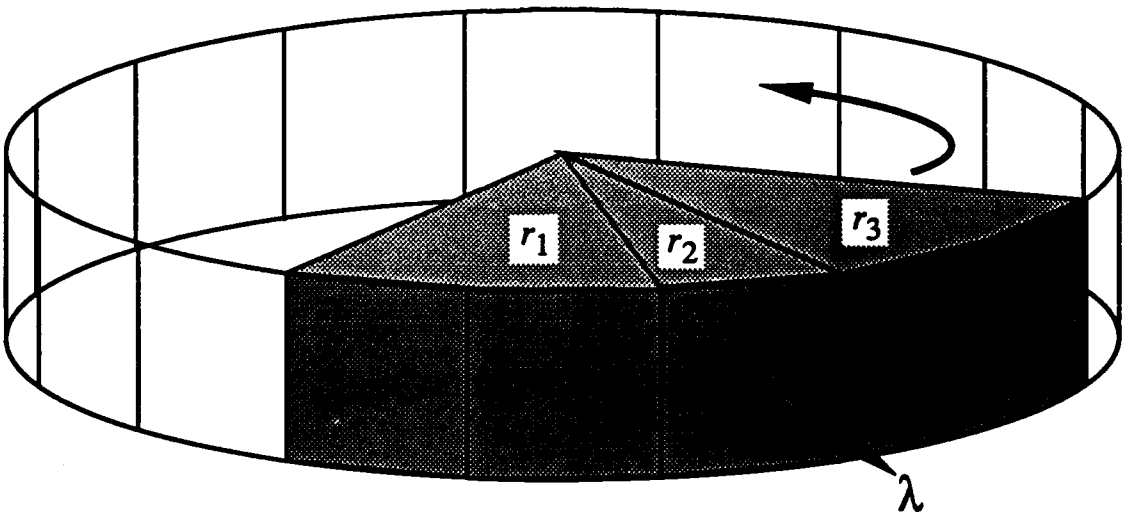


Figure 5.2.2-1. Arrangement of scene and sensor in one-dimensional optimum entropy minimization experiment.

each with some probability of being true and with entropy defined (in bits) as

$$E = -\sum_{i=1}^{N} P(H_i\ true)\log_2 P(H_i\ true) \tag{5-24}$$

Initially, $P(H_i\ true)=\frac{1}{N}$ for all $i=1...N$, and the hypothesis entropy is at a maximum value of $E=\log_2 N$ bits. The *a posteriori* probability on the states of nature is

$$P(x_i|R) = \frac{P(R|x_i)P(x_i)}{P(R)} = \frac{1}{N}\frac{P(R|x_i)}{P(R)} = \frac{1}{N}\frac{P(r_1, r_2, r_3|x_i)}{P(r_1, r_2, r_3)} \tag{5-25}$$

where r_1, r_2, and r_3 are the rexel values of the foveal sensor frame $R=\{r_1, r_2, r_3\}$. Because sensor noise effects are not considered in this experiment, the conditional probability of the rexel data from a registration given some state of nature x_i is 1 if the rexel values equal the sum of the corresponding scene pixels, and zero otherwise:

$$P(r_1, r_2, r_3|x_i) = \begin{cases} & x_i(\lambda-2)+x_i(\lambda-1)=r_1 \\ 1 & x_i(\lambda)=r_2 \\ & x_i(\lambda+1)+x_i(\lambda+2)=r_3 \\ 0 & \text{otherwise} \end{cases} \tag{5-26}$$

where λ is the foveal axis and $x_i(j)$ is the j'th pixel of the state of nature x_i. Since the scene is circular, $x_i(kN+j)=x_i(j)$ for any integers k and j. The *a priori* frame data probability is

$$P(r_1, r_2, r_3) = \sum_{\text{all } x_i} P(r_1, r_2, r_3|x_i)P(x_i) = \frac{1}{N}\sum_{\text{all } x_i}\left\{\begin{matrix} 1 & \text{if } R \text{ agrees with } x_i \\ 0 & \text{otherwise} \end{matrix}\right\} = \frac{c}{N} \tag{5-27}$$

where c is the number of states of nature agreeing with R. The *a posteriori* probability on the states of nature becomes

$$P(x_i|R) = \begin{cases} \frac{1}{c} & \text{scene pixels agree with rexels } r_1, r_2, \text{ and } r_3 \\ 0 & \text{scene pixels do not agree with rexels } r_1, r_2, \text{ and } r_3 \end{cases} \tag{5-28}$$

Given the rexel data from n foveations, the *a posteriori* probability on the states of nature is likewise given by

$$P(x_i|R_1,\ldots,R_n)=\begin{cases}\frac{1}{c} & \text{scene pixels agree with all rexels}\\ 0 & \text{scene pixels do not agree with all rexels}\end{cases} \quad (5\text{-}29)$$

where c is now the number of states of nature agreeing with all the frames.[17] As additional sensor frames are obtained, this "agreement" test becomes more stringent. In turn, the value of c decreases, or in the worst case, remains the same. The latter case occurs when a frame offers no new information (i.e., when a location is revisited). The value of c is thus a measure of perception ambiguity. When c=1, the scene is deterministically identified. Nevertheless, the task is not to resolve the scene but to localize a target. The ambiguity of importance to the foveation strategy is not that of the scene but of the task hypotheses.

It is seen that the process of foveation without sensor noise clamps to zero the probability of some candidate states of nature (hence they are no longer feasible states), while raising the single probability value on the remaining candidate states as c decreases. This is not to say that the hypothesis probabilities are equivalued or zero; the probability of the j'th hypothesis is

$$P(H_j\ true)=\frac{\text{number of candidate states of nature with target in pixel } j}{c} \quad (5\text{-}30)$$

where a candidate state of nature is now specifically defined as a combination of the scene pixel values (N-1 clutter values, 1 target value) with non-zero (1/c) *a posteriori* probability.

As was illustrated in the preceding section, foveations dramatically reduce the number of candidate states of nature. The number of possible N pixel scenes is

$$c_0 = \underbrace{2^{N-1}}_{\substack{\text{combinations}\\ \text{of } N\text{-1 binary}\\ \text{clutter pixels}}} \times \overbrace{N}^{\substack{\text{possible}\\ \text{target}\\ \text{locations}}} \quad (5\text{-}31)$$

The term $2^{(N-1)}$ corresponds to the combinations from N-1 clutter pixels which take on the value 0 or 1, and the N term corresponds to the N different locations where the target may

[17] In general, c is the number of states of nature agreeing with the measurements against which the *a-posteriori* statistic is conditional.

be located.[18] The information from just the value r_2 of the first frame reduces the number of candidate states of nature by more than a factor of 2 to $2^{(N-2)}(N-1)$ if r_2 is a clutter value, or by a factor N to $2^{(N-1)}$ if r_2 is the target value. Of course, if r_2 is the target value, then the task is over because the target is localized.

If the task were to resolve the entire scene, then the system would continue to foveate, employing not the target location hypothesis entropy but the entropy of the *a posteriori* probability on the states of nature. However, a foveal architecture would not be recommended for such a task because there is no localization of relevance in the scene; the entire scene must be resolved with maximum acuity, calling for uniform resolution vision.

Two versions of the hypothesis entropy minimization control strategy are considered: the myopic strategy, and a two step strategy. Both strategies share the same approach; the latter is implemented as a two-fold nested iteration of the myopic strategy. The perception is the list α of candidate states of nature. Every foveation provides three constraints (rexel values) on states of nature candidacy. These constraints $R=\{r_1, r_2, r_3\}$ are used to filter α and obtain a new smaller α list. The target hypothesis probabilities are obtained by taking the histogram of target location in all the filtered candidate states of nature. When the histogram has only one non-zero column, then all candidate state(s) of nature agree on target location (hypothesis entropy $E=0$), and the target is localized.

5.2.2.1 Myopic Entropy Reduction

The flowchart for the myopic algorithm is given in Figure 5.2.2.1-1. The foveal axis location minimizing the expected hypothesis entropy (upon performing the single foveation) is determined by first foveating on all candidate states of nature in α at each candidate foveal axis location λ. Candidate locations are those to which the fovea has not yet been directed. This provides $\overline{R}_{\lambda|x_i}$, the rexel data expected from a foveation to λ given that the state of nature is x_i. The list of candidate states of nature α is filtered by each set of expected data $\overline{R}_{\lambda|x_i}$, generating for each $\overline{R}_{\lambda|x_i}$ a list of expected candidate states of nature $\overline{\alpha}_{\lambda|x_i}$ satisfying

[18] Note that the target value does not factor into the calculation for state space size.

$$P\left(x_i \in \alpha \middle| \overline{R}_{\lambda|x_i}\right) \neq 0 \tag{5-32}$$

Since only the states in α are considered in forming $\overline{\alpha}_{\lambda|x_i}$ (as opposed to the entire state space), the information from all previous foveations are incorporated. The target location hypothesis probabilities are computed for each $\overline{\alpha}_{\lambda|x_i}$ by (5-10), (5-30), and

$$P\left(H_j \; true \middle| \overline{\alpha}_{\lambda|x_i}\right) = \frac{\text{number of candidate states of nature in } \overline{\alpha}_{\lambda|x_i} \text{ with target in pixel } j}{\text{number of candidate states of nature in } \overline{\alpha}_{\lambda|x_i}} \tag{5-33}$$

The conditional hypothesis entropy (5-11) resulting from a foveation to λ given x_i is

$$E_{\lambda|x_i} = -\sum_{j=1}^{N} P\left(H_j \; true \middle| \overline{\alpha}_{\lambda|x_i}\right) \log_2 P\left(H_j \; true \middle| \overline{\alpha}_{\lambda|x_i}\right) \tag{5-34}$$

The expected entropy resulting from a foveation to λ without any assumption on the state of nature (other than it is in α) is obtained by (5-12) and

$$E_\lambda = \sum_{x_i \in \alpha} E_{\lambda|x_i} P(x_i) = c^{-1} \sum_{x_i \in \alpha} E_{\lambda|x_i} \tag{5-35}$$

where c is the number of candidate states of nature in α. The next foveal axis is chosen as the one for which E_λ is minimum.

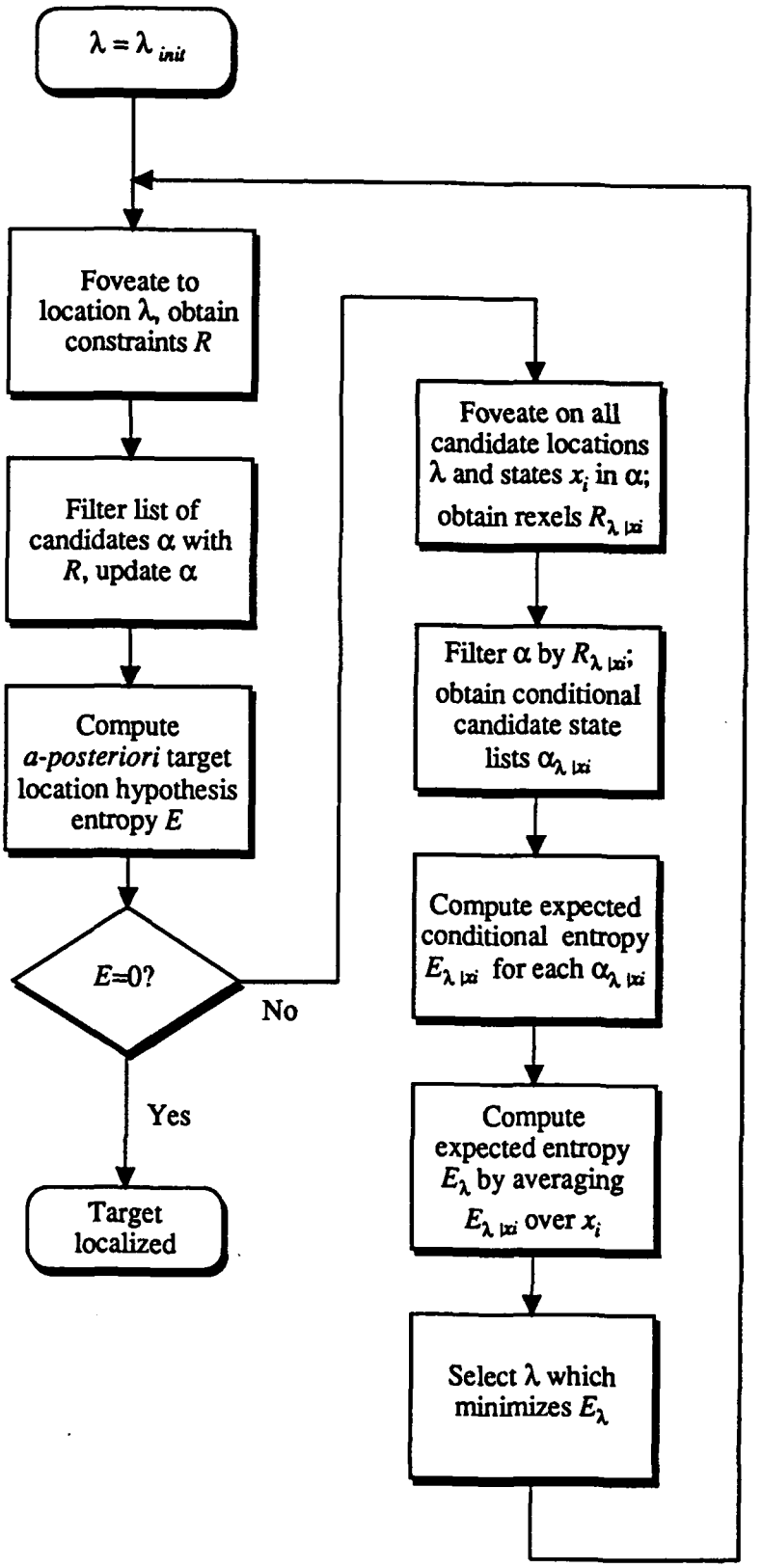


Figure 5.2.2.1-1. Algorithm flowchart for myopic entropy reduction gaze control.

5.2.2.2 Two Step Look-Ahead Entropy Reduction

The flowchart for the two step algorithm is given in Figure 5.2.2.2-1. Each candidate state of nature x_i in α is foveated at two candidate locations λ_m, λ_n. Recall that the order of locations in a sequence of foveations does not alter the final sequence entropy. The two step sequences $\{\lambda_m, \lambda_n\}$ considered can thus be limited to those satisfying

$$\lambda_m > \lambda_n \tag{5-36}$$

The list α is filtered by the resulting rexel data $\overline{R}_{\lambda_m|x_i}$ and $\overline{R}_{\lambda_n|x_i}$, generating the expected list of candidate states of nature $\overline{\alpha}_{\lambda_m,\lambda_n|x_i}$. The target location hypothesis probabilities are computed for each $\overline{\alpha}_{\lambda_m,\lambda_n|x_i}$ by

$$P\left(H_j \; true \middle| \overline{\alpha}_{\lambda_m,\lambda_n|x_i}\right) = \frac{\text{number of candidate states of nature in } \overline{\alpha}_{\lambda_m,\lambda_n|x_i} \text{ with target in pixel } j}{\text{number of candidate states of nature in } \overline{\alpha}_{\lambda_m,\lambda_n|x_i}} \tag{5-37}$$

The hypothesis entropy resulting from the foveation sequence $\{\lambda_m, \lambda_n\}$ given x_i is

$$E_{\lambda_m,\lambda_n|x_i} = -\sum_{j=1}^{N} P\left(H_j \; true \middle| \overline{\alpha}_{\lambda_m,\lambda_n|x_i}\right) \log_2 P\left(H_j \; true \middle| \overline{\alpha}_{\lambda_m,\lambda_n|x_i}\right) \tag{5-38}$$

The expected entropy resulting from the sequence $\{\lambda_m, \lambda_n\}$ without any assumption on the state of nature (other than it is inα) is obtained by

$$E_{\lambda_m,\lambda_n} = \sum_{x_i \in \alpha} E_{\lambda_m,\lambda_n|x_i} P(x_i) = c^{-1} \sum_{x_i \in \alpha} E_{\lambda_m,\lambda_n|x_i} \tag{5-39}$$

The sequence chosen $\{\lambda_m^*, \lambda_n^*\}$ is the one for which E_{λ_m,λ_n} is minimum.

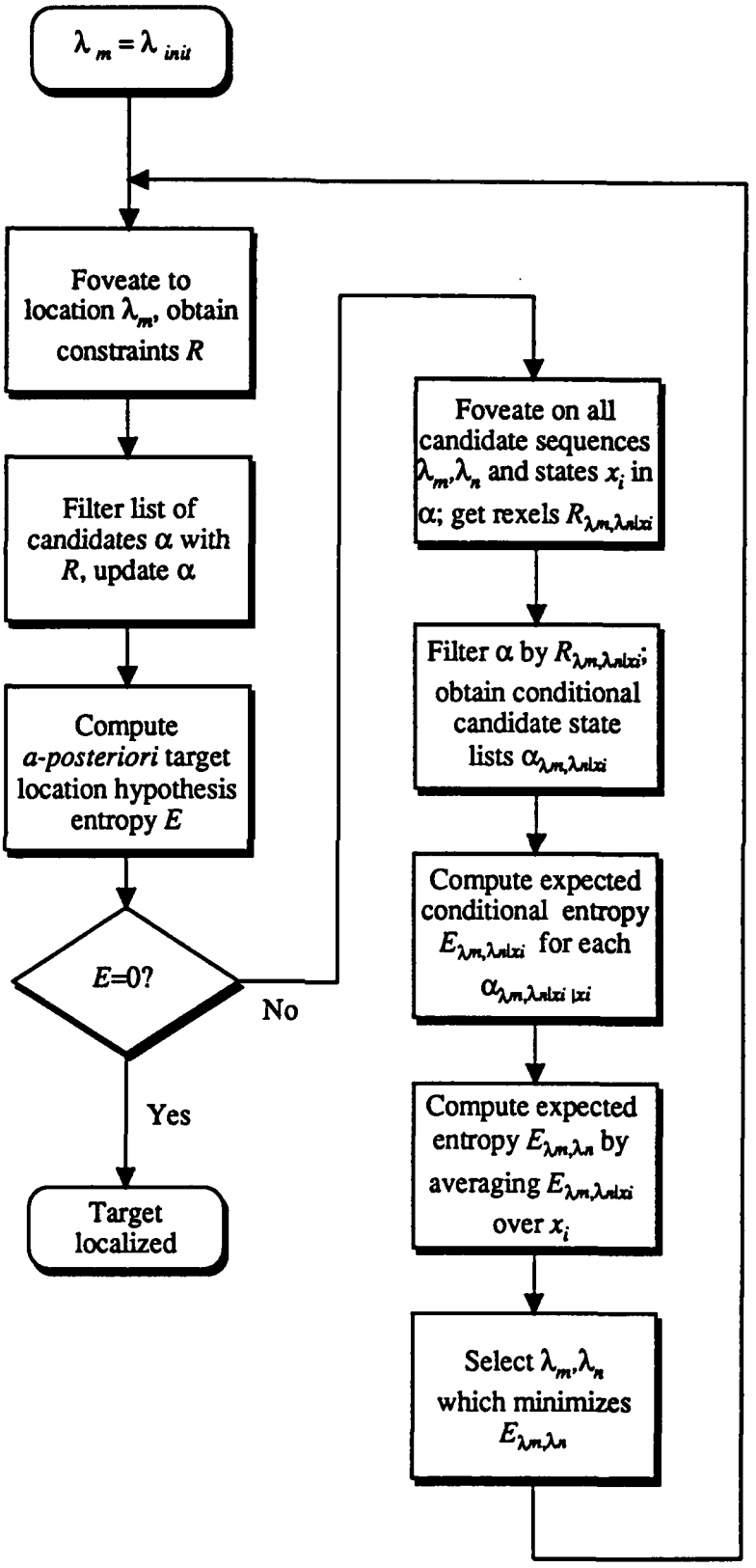


Figure 5.2.2.2-1. Algorithm flowchart for two step entropy reduction gaze control.

The two step strategy is repeated after every foveation. This is because $\{\lambda_m^*, \lambda_n^*\}$ represents the optimum strategy given the data obtained so far, which is itself represented by α, but may not be the optimum strategy given different data (a different α); if the foveation to λ_m^* is performed first, the new two-step optimum sequence may not include λ_n^* at all. The second location of a sequence may not be used due to its suboptimality given the data collected from the foveation to the first location. While the ordering of the two locations in the optimum sequence does not alter the expected hypothesis entropy, it does affect the actual system performance. The α list resulting from foveating to λ_m^* will in all likelihood be different from that resulting from a foveation to λ_n^*. The location providing the lowest expected myopic hypothesis entropy is used first, since the other location may be replaced by a different first location from the next computed sequence.

5.2.2.3 *N*-Step Look-Ahead Entropy Reduction

The two step strategy is extended into an n step strategy by considering all sequences (combinations) of n candidate locations. As with the two step strategy, the n step sequences $\{\lambda_{m_1}, \lambda_{m_2}, \ldots, \lambda_{m_n}\}$ considered can be limited to those with candidate locations satisfying

$$\lambda_{m_1} > \lambda_{m_2} > \ldots > \lambda_{m_n} \tag{5-40}$$

Of course, the longer the sequence, the greater is the number of sequence combinations. Consequently, more conditional expected lists of candidate states of nature $\overline{\alpha}_{\lambda_1,\lambda_2,\ldots,\lambda_n|z_i}$ and corresponding hypothesis entropies $E_{\lambda_1,\lambda_2,\ldots,\lambda_n|z_i}$ must be generated, and the search space for the minimum expected hypothesis entropy $E_{\lambda_1,\lambda_2,\ldots,\lambda_n}$ is larger.

5.2.2.4 Numerical Results

A foveal machine vision system for unresolved target localization in the described 1-D scene was simulated using the algorithms for myopic and two step look-ahead entropy. The minimum target pixel value of 2 was used to to provide a worst case target-to-background signal ratio. A ten pixel wrap-around scene was used with the values as shown in Figure 5.2.2.4-1.

scene address:	1	2	3	4	5	6	7	8	9	10
scene value:	0	0	0	1	1	1	0	0	0	2

Figure 5.2.2.4-1. Test 1-D scene.

The rexels $R=\{r_1, r_2, r_3\}$ are labeled as highlighted in Figure 5.2.2.4-2. In this illustration, the foveal axis (r_2) is directed to location 5.

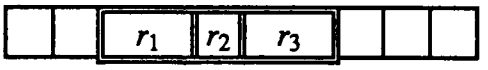


Figure 5.2.2.4-2. Test rexel coverage.

Scene pixel 4 is selected as the initial foveal axis (Figure 5.2.2.4-3), and the rexel data from the first registration is $R=\{0,1,2\}$. The rexels of the first registration resolve some scene locations sufficiently to discard any possibility of target presence ("X"), but other locations are not resolved sufficiently or at all ("?"). This combination of scene pixel values and initial axis location produces a relatively high ambiguity in target location.

scene address:	1	2	3	4	5	6	7	8	9	10
rexels R_1:	0	0	0	1	1	1	0	0	0	2
perception:	?	X	X	X	?	?	?	?	?	?

Figure 5.2.2.4-3. Initial registration.

The list of candidate states of nature α satisfying the constraints of R contains 288 10-tuple vectors. The target location entropy within these states of nature is 2.7255 bits. This entropy indicates that target location is resolved to $2^{2.7255} \approx 6.6$ locations. This is less than the seven question marks illustrated above because of the partial knowledge on the values of locations 5 and 6.

The system proceeds to determine the next foveal axis. The myopic strategy is considered first (Figure 5.2.2.1-1). The expected target location entropy after foveating to each candidate location in the candidate states of nature in α is given in Table 5.2.2.4-1.

Second Foveal Axis Location	Expected Hypothesis Entropy (bits)
1	1.77
2	2.25
3	0.89
5	0.89
6	0.94
7	1.77
8	0.94
9	1.35
10	1.32

Table 5.2.2.4-1. Expected myopic hypothesis entropy upon initial registration.

Since sensor noise is considered negligible, a repeated foveation to location 4 would not produce any new or different information, and the hypothesis entropy would remain at 2.7255. The myopic strategy dictates locations 3 or 4 as the next foveal axis locations because they offer the minimum hypothesis entropy after the single foveation. Location 3 is arbitrarily selected.

The foveation to location 3 (Figure 5.2.2.4-4) produces the rexel data $R=\{0,0,2\}$. The α list contains 64 candidate states of nature and a hypothesis entropy of 2 bits; target location is narrowed down to the last four locations in the scene.

scene address:	1	2	3	4	5	6	7	8	9	10
rexels R_1:	0	0	0	1	1	1	0	0	0	2
rexels R_2:	0	0	0	1	1	1	0	0	0	2
perception:	X	X	X	X	X	X	?	?	?	?

Figure 5.2.2.4-4. Second myopic registration.

The hypothesis entropy expected after foveating to the different candidate locations in the α list vectors is given in Table 5.2.2.4-2. It is interesting to note the statistical symmetry about the scene when bisected into the regions {4,5,6,7,8} and {9,10,1,2,3} (recall that the environment is circular, and that the expected entropy from a foveation to location 3 or 4 is 2.0 bits).

Second Foveal Axis Location	Expected Hypothesis Entropy (bits)
1	1.34
2	1.19
5	1.19
6	1.34
7	0.73
8	0.50
9	0.50
10	0.73

Table 5.2.2.4-2. Expected myopic hypothesis entropy upon second registration.

It is also interesting to note that the entropy expected after foveating to a location can be lower in an earlier iteration. For example, after the first registration, a foveation to location 5 was expected to produce a hypothesis entropy of 0.89. After the current (second) registration, a foveation to location 5 is expected to produce a hypothesis entropy of 1.19. This is because the statistics of target location in the state vectors in the α list change as the list is constrained by new measurements (rexel data). A poorly resolved region could contain the target and thus a foveation there provides significant information (confirming or denying target presence), whereas foveating to a well resolved region does not introduce as much new information. A foveation to location 5 was earlier believed to

possibly reduce the hypothesis entropy. However, the foveation to location 3 precluded any target presence in the region {5, 6}, and the expected information (entropy reducing capacity) of a foveation to location 5 is sharply reduced (current entropy is not expected to significantly change). The distribution of ambiguity changes as foveations are performed, but the total ambiguity never increases.

Locations 8 and 9 produce the minimum expected entropy, and location 8 is selected. The foveation to location 8 produces the rexel data R={1,0,2} (Figure 5.2.2.4-5). The α list contains 4 candidate states of nature and a hypothesis entropy of 1 bit; target location is narrowed down to the last two locations in the scene.

scene address:	1	2	3	4	5	6	7	8	9	10
rexels R_1:	0	0	0	1	1	1	0	0	0	2
rexels R_2:	0	0	0	1	1	1	0	0	0	2
rexels R_3:	0	0	0	1	1	1	0	0	0	2
perception:	X	X	X	X	X	X	X	X	?	?

Figure 5.2.2.4-5. Third myopic registration.

The hypothesis entropy expected after foveating to the remaining candidate locations in the α list vectors is given in Table 5.2.2.4-3. Some locations offer no new information (entropy remains at 1 bit). Other locations offer an expected entropy of zero, meaning that foveating to these locations will deterministically localize the target. A non-zero expected entropy for a foveation to some location is only a statistical measure which may or may be the actual entropy after the foveation.[19] On the other hand, a zero valued expected entropy is a deterministic guarantee of target localization because in the computation of that value, all candidate states of nature are considered. Zero entropy values are possible only because sensor noise is considered negligible in this exercise; otherwise, it would introduce a certain degree of ambiguity (entropy), reducible only through repeated sampling and averaging to reduce perception variance.[20]

---

[19] If the expected entropy for a candidate location equals the current hypothesis entropy, then it is also known deterministically that such a foveation offers no new information.

[20] Measurable sensor noise would also not permit the collapse of the α list, since under unbounded (e.g., Gaussian) noise, all states of nature would have some finite probability of existing.

Second Foveal Axis Location	Expected Hypothesis Entropy (bits)
1	1.00
2	0.00
5	1.00
6	1.00
7	0.00
9	0.00
10	0.00

Table 5.2.2.4-3. Expected myopic hypothesis entropy upon third registration.

The next foveation is directed at location 2 (Figure 5.2.2.4-6), producing the rexel data $R=\{2,0,1\}$. The α list contains 2 candidate states of nature, but since the target is at location 10 in both state vectors, the hypothesis entropy is 0; the target is localized to the maximum resolution of the system at the location labeled "!" in the perception. Four registrations were required to complete the task.

scene address:	1	2	3	4	5	6	7	8	9	10
rexels R_1:	0	0	0	1	1	1	0	0	0	2
rexels R_2:	0	0	0	1	1	1	0	0	0	2
rexels R_3:	0	0	0	1	1	1	0	0	0	2
rexels R_4:	0	0	0	1	1	1	0	0	0	2
perception:	X	X	X	X	X	X	X	X	X	!

Figure 5.2.2.4-6. Fourth myopic registration.

The two step strategy is now considered (Figure 5.2.2.1-2). After the initial registration to location 4, the expected entropy resulting from foveations to all the combinations of candidate location pairs are computed (Table 5.2.2.4-4). The entropy values are given in Table 5.2.2.4-4. The entropy for sequences with $\lambda_1<\lambda_2$ is the same as that for the sequences with the locations reversed. Also, the entropy for the sequence $\{\lambda_i, \lambda_i\}$ is the same as that for a single foveation to λ_i.

1	1.77								
2	1.28	2.25							
3	0.60	0.53	0.89						
5	0.32	0.60	0.53	0.89					
6	0.50	0.68	0.60	0.53	0.94				
7	0.79	1.35	0.32	0.60	0.22	1.77			
8	0.60	0.50	0.22	0.32	0.32	**0.00**	0.94		
9	0.89	1.25	0.22	0.22	0.51	0.89	**0.00**	1.35	
10	0.85	1.32	0.32	0.22	0.11	0.89	0.22	0.44	1.32
λ_1/λ_2	1	2	3	5	6	7	8	9	10

Table 5.2.2.4-4. Expected two step hypothesis entropy upon initial registration.

Four sequences are expected to achieve target localization: {8,7}, {9,8}, and their conjugates {7,8}, {8,9}. Of these, the sequence selected is that which minimizes the entropy after the first foveation. The expected hypothesis entropies after foveating to locations 7, 8, and 9 are 1.77, 0.94, and 1.35 respectively. Thus, the sequences {8,7} and {8,9} are the optimum two step choices, and the first is arbitrarily selected. Note that expected entropy after the first foveation is greater than that of foveating to locations 3 or 4, which are the myopic choices. Thus, the two foveation strategies commence immediately in different directions.

The foveation to location 8 produces the rexel data R={1,0,2} (Figure 5.2.2.4-7). The α list is reduced to 14 candidate states of nature and the hypothesis entropy is 1.95 bits. The ambiguity of $2^{1.95}$ is less than five, the number of interrogation marks in the perception, because of the partial information on locations 5,7,9, and 10. It is interesting to observe that this entropy is less than that obtained after a foveation to location 3 under the myopic strategy, even though the entropy was expected to be greater. This is purely a consequence of the true state of nature.

scene address:	1	2	3	4	5	6	7	8	9	10
rexels R_1:	0	0	0	1	1	1	0	0	0	2
rexels R_2:	0	0	0	1	1	1	0	0	0	2
perception:	?	X	X	X	?	X	?	X	?	?

Figure 5.2.2.4-7. Second two step registration.

The expected hypothesis entropies are computed for the sequences of candidate location pairs. These values are given Table 5.2.2.4-5 in units of bits.

1	1.36							
2	**0.00**	0.86						
3	0.57	**0.00**	0.57					
5	0.57	0.29	0.57	1.09				
6	0.57	0.29	0.57	1.09	1.09			
7	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**		
9	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	
10	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
λ_1/λ_2	1	2	3	5	6	7	9	10

Table 5.2.2.4-5. Expected two step hypothesis entropy upon second registration.

Several sequences guarantee target localization with just one more foveation. The sequence {7,1} is chosen arbitrarily. Foveating to location 7 produces the rexel data R={2,0,0} (Figure 5.2.2.4-8). The resulting α list contains four candidate, each with the target at location 10. Hypothesis entropy is zero (as guaranteed), and the target is localized.

scene address:	1	2	3	4	5	6	7	8	9	10
rexels R_1:	0	0	0	1	1	1	0	0	0	2
rexels R_2:	0	0	0	1	1	1	0	0	0	2
rexels R_3:	0	0	0	1	1	1	0	0	0	2
perception:	X	X	X	X	X	X	X	X	X	!

Figure 5.2.2.4-8. Third two step registration.

The two step strategy was able to localize the target with fewer registrations than the myopic strategy (in this small example, three registrations versus four). The computational overhead incurred by considering the second location in the expected entropy calculations was negligible. This is because the conditional list of candidate states of nature after the first proposed foveation $\overline{\alpha}_{\lambda_1,\lambda_2,\ldots,\lambda_n|x_i}$ is much smaller than the original α list, as is evident by observing the rapid collapse of the α list with each foveation. The second step statistics

are thus computed quickly. The performance of the two strategies in this exercise are summarized in Tables 5.2.2.4-6 and 5.2.2.4-7. Note that the α list of both approaches contains four candidate states after the third registration, yet the two step list entropy is zero.

Iteration	Foveal Axis Location	Candidate States	Hypothesis Entropy
1	4	288	2.73
2	3	64	2.00
3	8	4	1.00
4	2	2	0.00

Table 5.2.2.4-6. Summary of myopic strategy performance.

Iteration	Foveal Axis Location	Candidate States	Hypothesis Entropy
1	4	288	2.73
2	8	14	1.95
3	7	4	0.00

Table 5.2.2.4-7. Summary of two step strategy performance.

5.2.3 Approximation to Entropy Minimization

The foveation strategies employing the strict minimization of hypothesis entropy and reversible integrated perceptions are not computationally tractable for scenarios of any significant size.[21] However, the tractability of foveation strategies improves greatly when the assumptions which led to the approximate state of nature integrated perception in Section 4.4 are employed.

Consider the discard method approach to the integrated perception introduced in Section 4.5. Its underlying assumption is the independence of pixel statistics from frame to frame of rexel data. If the linear minimum mean square estimate (LMMSE) of a pixel, given by equations (4-33) and (4-34), is employed, then the additional assumption of Gaussian distribution of scene pixels and noise is introduced. This approach to the

[21] A DEC VAX 11/780 was never able to complete the selection of the second foveal axis within a 15 pixel 1-D scenario in the 23 hours of system availability between system back-ups (at which time all active jobs are cancelled).

maintenance of the integrated perception can be considered as a single step Bayesian learning process. The discard of lower acuity information for new higher acuity information is analogous to "forgetting" the first lesson as a more informative second lesson is learned.

The issue remains of where to direct the sensor optical axis. Entropy is a integrated measure of ambiguity across all hypotheses, or specifically, the weighted sum of the self-information of each event hypothesis $\mathfrak{R}$ given the measured data $\mathfrak{I}$

$$-P(\mathfrak{R}|\mathfrak{I})\log_2 P(\mathfrak{R}|\mathfrak{I}) \qquad (5\text{-}41)$$

In some tasks, such as localizing an unresolved target, there is a direct correlation between a hypothesis and a location within the field-of-regard. Weighted self-information can thus be treated as a spatially localized measure of perception ambiguity. One possible heuristic approach to foveation control is to allocate acuity where the weighted self-information is highest. The next optical axis is selected such that it maximizes the convolution of the self-information function throughout the field-of-regard with the acuity profile of the foveal sensor (Figure 5.2.3-1).

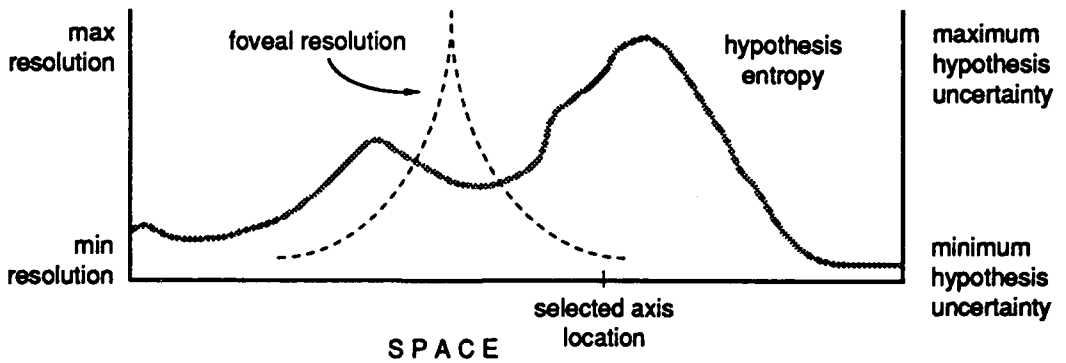


Figure 5.2.3-1. Foveation by convolution of foveal resolution with hypothesis entropy.

The selection of initial saccades in the human visual system seems to follow a similar algorithm. When there is no prior knowledge given on the location of targets and nontargets, initial saccades are directed towards the centroid of cue clusters. The Bayesian strategy for gaze control can take advantage of any available *a priori* information to fine tune saccades. This mechanism also exists in human vision where given some prior information on target location, initial saccades are biased away from the cue cluster centroid

and towards the given target location [He89]. This implies that reflexive low level behavior such as the initial saccade upon scene projection is connected to an integrated scene perception; in this case, the initial state of the perception contains the target probability verbally provided by the experiment conductor.

In some applications, presupposed scene activity provides implicit prior knowledge of target location. For example, the early detection and localization of very distant airborne targets by a ground based system may be accelerated if the field-of-regard just above the horizon is emphasized. If scene dynamics are not expected to alter significantly from the presupposed activity , this information can be "hard-wired" into the vision system. This is observed in the human visual system, where horizontal target motion tracking (i.e., tracking a running animal) is significantly more accurate than vertical motion tracking [Baloh88].

5.3 Likelihood Ratio Maximization

A cost function and resulting strategy which contrasts with entropy minimization is the likelihood ratio. This is the cost function of the interrogation mode control strategy. The likelihood ratio $\Lambda_{x,\Re}$ of an observation x of a hypothesis $\Re$ is

$$\Lambda_{x,\Re} = \frac{P(x|\Re)}{P(x|\overline{\Re})} \tag{5-42}$$

where $\overline{\Re}$ is the null hypothesis

$$P(\Re \cup \overline{\Re}) = 1 \tag{5-43}$$

$$P(\Re \cap \overline{\Re}) = 0 \tag{5-44}$$

A likelihood ratio $\Lambda_{x,\Re} > 1$ indicates that the observation is supported by the hypothesis. A likelihood ratio $\Lambda_{x,\Re} < 1$ indicates that the observation contradicts the hypothesis. This approach is straightforward in the case of many hypotheses when each hypothesis $\Re_i$ can be tested independently with a corresponding null hypothesis $\overline{\Re}_i$.

The foveation strategy resulting from the likelihood ratio cost function acquires observations in an attempt to confirm or deny a particular hypothesis (as opposed to global learning). The computed saccades are aimed directly at the location in the scene associated with the hypothesis. Thresholds can be set to establish desired confidence intervals and type 1 classification error, which is defined as rejecting the null hypothesis $\overline{\mathfrak{R}}$ when it is true [Larson74]. It has been proposed that the human visual system may perform an approximation to likelihood ratio maximization when locating signals in the field-of-view [Caelli87].

Since a particular hypothesis is being tested, the foveation strategy interrogates only the region in the scene associated with the hypothesis. Consequently, the likelihood ratio cost function results in localized foveations as opposed to the global learning foveations of entropy minimization. As with entropy minimizations, perceptions representing approximations to the state of nature can be employed for computational tractability.

5.4 Control Strategies for Target Localization

A simple control strategy for target localization was presented in Section 3.5 to illustrate the relationship between foveal geometry and machine vision system performance. The strategy assumed a single unresolved target against negligible clutter and sensor noise, and consisted of foveating to the center of the brightest rexel in the last frame. The task of unresolved target localization is now revisited using the techniques described in Chapters 4 and 5. Multiple targets, background clutter, and sensor noise will be introduced into the analysis. The discard method of perception generation will be employed. The perception cues consist of relatively bright unisource regions which can be generated by a bright target or by localized clutter. Low acuity sensing provides these cues but cannot distinguish one source from the other on the basis of region brightness alone. The sources are resolved by interrogating the cues.

5.4.1 Top Level System Operation

There are N^2 hypotheses, each proposing the existence of an unresolved target in a unique pixel. As proposed in Section 5.1, target localization is segmented into a survey mode, which efficiently searches for cues exploiting the wide field-of-view of low acuity peripheral vision, and an interrogation mode which resolves the cues. The survey mode corresponds to a surveillance operation, while the interrogation mode tests the hypothesis of a particular target state. By permitting more than one target to appear in the scene, independence of hypothesis testing is achieved; each hypothesis has a single and independent null hypothesis. Also, a hypothesis entropy can be defined at each scene pixel.[22]

An integrated low level perception is maintained containing the LMMSE estimates and estimate variance of pixel values from rexel data. A second higher level perception is maintained, which is directly generated from the first. It contains the *a posteriori* probability of a target existing at each pixel in the field-of-regard, given the pixel value estimates. The system's default mode of operation is the survey mode, in which it tries to learn about the entire scene as a whole. The array of hypothesis entropy values is convolved with the acuity profile of the foveal sensor to select an optical axis which is expected, upon foveation, to minimize the entropy of the entire perception.

After each entropy minimizing foveation, the low level and high level perceptions are updated, and any cues are evaluated. A cue is defined as a unisource region in the high level perception indicating high probability of target existence. When a cue exceeding a threshold probability is detected, the system mode switches from survey to interrogation. In this second mode, the system foveates to the centroid of the unisource region measuring the cue. In the case of multiple cues, the system foveates to the centroid of the unisource region with maximum target probability. All interrogation mode foveations are directed within the boundary of the unisource region. The perception of the cue (the unisource region) is resolved into smaller unisource regions to test the hypothesis of target existence. The system remains in this mode, resolving the cue by foveating to the smaller and most

[22] In the case of multiple targets, the presence of a target at a pixel is associated with a unique binary random variable, and there are N^2 random variables altogether. The entropy of each scene pixel is that of the *a-posteriori* probability distribution function of the associated random variable. In the case of a single target, all hypotheses and pixels are associated with a single random variable which takes on one of N^2 values. In the latter case, only a self-information term is associated with each scene pixel, as opposed to a formal entropy value in the former case.

promising unisource regions within the foveation boundaries, until the hypothesis is confirmed or denied with sufficient confidence. The system then reverts to the survey mode. The overall search terminates when the probability of one or more unlocalized targets drops below an acceptable false negative threshold.

5.4.2 Survey Mode

The objective of the survey mode is to learn about target presence throughout the field-of-regard. Since the survey mode corresponds to a general learning, overall hypothesis entropy is an appropriate figure of merit, and the strategy employed in this mode is that of entropy minimization.

The low level perception accounts for clutter and sensor noise. The clutter and noise statistics are assumed to be known. The larger the rexel, the more clutter is averaged into the rexel value. In the noise-free ($\sigma_n^2=0$) case, the LMMSE value $\hat{x}$ is the average pixel value which generates the rexel measurement, and the variance of the pixel estimate value (error variance) e_m is the clutter variance σ_c^2. Through the Δ term, sensor noise has the effect of discrediting the measurement data biasing the pixel estimate away from the measurement term $\frac{v_{r,m}}{m}$ and closer to the *a priori* clutter mean term μ_x.

The high level perception is the *a posteriori* probability given $\hat{x}$ and e_m of the target being present in a pixel within the field-of-regard. Let the *a priori* knowledge include, in addition to the clutter and noise statistics, the first and second order statistics μ_t and σ_t^2 of the target luminosity v_t. Also required is an estimate for the number of targets n_t in the field-of-regard. The *a priori* probability of a target being at a particular pixel in a field-of-regard consisting of $N \times N$ pixels is simply

$$P_{t,x,y}(t=1) = \frac{n_t}{N^2} \tag{5-45}$$

where t is the binary random variable representing target present (t=1) and target absent (t=0), $n_t<N^2$, and it is assumed that no more than one target appears at any given pixel.[23]

[23] If $n_t<<N^2$, the probability of multiple targets appearing at a given pixel is negligible.

The distribution of the estimated pixel value given the rexel value $v_{r,m}$ is a Gaussian function with mean $\hat{x}$ and variance e_m:

$$p_x(v) = N(v|\hat{x}, e_m) \tag{5-46}$$

where $N(a|b,c)$ represents a Gaussian distribution of the random variable a with mean b and variance c. The scene luminosity l at the pixel is the value of the clutter at the pixel location if no target is present, or the sum of clutter and target luminosity if a target is present. The corresponding *a priori* probability distributions are

$$p_l(l|t=0) = N(l|\mu_c, \sigma_c^2) \tag{5-47}$$

and

$$p_l(l|t=1) = N(l|\mu_c + \mu_t, \sigma_c^2 + \sigma_t^2) \tag{5-48}$$

respectively, the latter being the convolution of the clutter and target *a priori* distributions. The probability of obtaining the estimate pixel value $\hat{x}$ in the data set $D=(\hat{x}, e_m)$ is the integral of the data set distribution weighted by the distribution of the scene luminosity:

$$p_D(D|t=0) = \int_{-\infty}^{\infty} p_x(l) p_l(l|t=0) dl \tag{5-49}$$

$$p_D(D|t=1) = \int_{-\infty}^{\infty} p_x(l) p_l(l|t=1) dl \tag{5-50}$$

The term D is a formal random variable with a probability distribution derived from the convolution of two Gaussian distributions [Papou84]:

$$\begin{aligned}
\int_{-\infty}^{\infty} p_D(D|t)\, d\hat{x} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_x(l) p_l(l|t)\, dl d\hat{x} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} N(l|\hat{x}, e_m) N(l|\mu_c + t\mu_t, \sigma_c^2 + t\sigma_t^2)\, dl d\hat{x} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} N(l-\hat{x}|0, e_m) N(l|\mu_c + t\mu_t, \sigma_c^2 + t\sigma_t^2)\, dl d\hat{x} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} N(\hat{x}-l|0, e_m) N(l|\mu_c + t\mu_t, \sigma_c^2 + t\sigma_t^2)\, dl d\hat{x} = 1
\end{aligned} \tag{5-51}$$

Thus, if the pixel value were known unambiguously (no sensor noise, m=1, and e_m=0), the probability of obtaining data set D would be

$$p_D(D|t=0)\Big|_{e_m=0} = \int_{l=0}^{\infty} \delta(l-\hat{x}) p_t(l|t=0)dl = p_t(\hat{x}|t=0) \tag{5-52}$$

$$p_D(D|t=1)\Big|_{e_m=0} = \int_{l=0}^{\infty} \delta(l-\hat{x}) p_t(l|t=1)dl = p_t(\hat{x}|t=1) \tag{5-53}$$

where $\delta(x)$ is the dirac function.

The *a posteriori* probability of a target being at a particular pixel is obtained through Bayes rule:

$$P_t(h|D) = \frac{p_D(D|t=h)P_t(h)}{p_D(D)} = \frac{p_D(D|t=h)P_t(h)}{\sum_{t=0}^{1} p_D(D|t)P_t(t)}$$

$$= \frac{\left(\frac{n_t}{N^2}\right)^h \left(1-\frac{n_t}{N^2}\right)^{1-h} \int_{l=0}^{\infty} p_p(l) p_t(l|t=h)dl}{\left(1-\frac{n_t}{N^2}\right)\int_{l=0}^{\infty} p_p(l) p_t(l|t=0)dl + \left(\frac{n_t}{N^2}\right)\int_{l=0}^{\infty} p_p(l) p_t(l|t=1)dl} \tag{5-54}$$

where h is the value of the hypothesis being tested (h=1 is target present, h=0 is target absent). The value of $P_t(h|D)$ for every pixel in the field-of-regard forms the high level perception. The entropy of the hypothesis on each pixel is given by

$$E = -\sum_{t=0}^{1} P_t(t|D)\log_2\left[P_t(t|D)\right] \tag{5-55}$$

The optical axis location selected for the next foveation is the one which maximizes the convolution of high level perception entropy with foveal sensor acuity

$$C(x,y) = \sum_{i=x_{min}}^{x_{max}} \sum_{j=y_{min}}^{y_{max}} \frac{E_{i,j}}{m(x-i,y-j)} \tag{5-56}$$

where $E_{i,j}$ is the entropy of the target existence hypothesis at the pixel in location (i,j), $m(x,y)$ is the linear dimensions of the rexel at location (x,y) relative to the geometry center, and (x_{min}, x_{max}) and (y_{min}, y_{max}) are the limits on the field-of-regard.

5.4.3 Interrogation Mode

The objective of the interrogation mode is to confirm or deny (to within specified false positive or false negative probability thresholds) a specific hypothesis (i.e., cue) which is particularly strong. System attention focuses on that particular cue until the interrogation is complete (several stop rules are presented). Since the interrogation mode corresponds to the maximization of a particular hypothesis probability, the strategy employed is that of likelihood maximization.

In addition to computing the hypothesis entropy for each pixel, the hypothesis likelihood ratio Λ_t is also computed:

$$\Lambda_t = \frac{P_t(D|t=1)}{P_t(D|t=0)} \tag{5-57}$$

The foveation controller switches from survey mode to interrogation mode when a unisource region of the perception has an average likelihood ratio greater than some threshold Λ_{switch}. When this condition is met, the foveation controller orients the optical axis to the center of the unisource region with maximum likelihood ratio. Note that since all the perception information in a unisource region originates from the same source (a rexel), the likelihood ratio value of the pixels in this region will be the same.

Upon each generation of a foveal sensor frame, the low and high level perceptions are updated, and the original unisource region under interrogation is resolved into smaller unisource regions. In the interrogation mode, the optical axis is repositioned every time to the new unisource region with maximum likelihood ratio which falls within the boundaries of the initial unisource region. The system thus interrogates only the most likely hypothesis that initially triggered the mode change.

The machine vision system continues to resolve the original unisource region, guided by maximum likelihood ratio, until one of three stop rules are encountered: hypothesis confirmed, hypothesis rejected, or hypothesis unresolved. Each stop rule is controlled by a predetermined threshold selected to obtain the desired confidence interval and maximum dwell time.

The hypothesis of target present at a pixel is confirmed when Λ_i for a singly resolved pixel in original unisource region is greater than a predetermined threshold $\Lambda_{confirm}$. This threshold can be determined by maximizing the probability of true detection while keeping the probability of false alarm below some specified value. When this occurs, the pixel is flagged as true and the system returns to its default survey mode. If the cue was caused by a cluster of clutter, then upon resolving the unisource region the values for all likelihood ratios in the region decrease. When the overall probability of one or more undiscovered targets in the interrogation region drops below a threshold, the region if flagged as false and the system returns to its default survey mode. This threshold is determined in part by the permissible false negative system requirements.

If the system is in the interrogation mode for a maximum foveation limit F_{max} without meeting any of the two previous stop rules, as can occur particularly in the case of high sensor noise which encourages multiple overlapping foveations for temporal filtering, the unisource region is flagged as indeterminate and the system returns to its default survey mode. To avoid overinterrogating small regions or underinterrogating large regions, F_{max} can be proportional to the area of the original unisource region. In this case, F_{max} is interpreted as the maximum number of interrogating foveations permitted per unit area of field-of-regard.

Multiple cues can appear with likelihood ratios above Λ_{switch} from a single survey mode foveation. Additional cues with likelihood ratios above Λ_{switch} may appear in the perception as a result of interrogation mode foveations. Upon completion of an interrogation procedure (completion defined by the stop rules), the system may immediately commence the interrogation of a different cue. The flagging of pixels and regions when leaving an interrogation procedure prevents infinite loops (system attempting to interrogate a recently interrogated cue).

5.5 Additional Remarks

The gaze control strategies presented in this chapter produce sensor movement analogous to saccades in biological eye movement. Experiments have shown that human eye movement in tasks involving static scene consists of the interrogation of strong cues

and survey-like excursions across regions in the field-of-regard not interrogated or previously surveyed [Yarbus67]. The implementation of foveal sensor gaze control strategies producing sensor movement analogous to smooth pursuit eye movement remains to be addressed. Such strategies can employ existing algorithms for target tracking with uniform resolution sensors, since they share the common objective of keeping the target centered in the field-of-view.

The form of the hypothesis probability as a function of the state of nature may not be readily available. For example, a machine vision system inspecting parts on an assembly line for defects can conceivably encounter an unlimited range of variations to some reference labeled as a "good" part. Even this reference may include a range of tolerances. The two hypotheses, "part is good" or "part is bad", are of little assistance to the task of foveation. Defining each hypothesis as the presence of a variant of a part (good or defective) results in an unwieldingly large number of hypotheses, although the relationship of each hypothesis to the state of nature can be clear and thus be of value to the selection of a new gaze angle. Hierarchical hypotheses can be used with hierarchical object models and the different (higher and lower level) perceptions generated by a machine vision system.

Chapter 6

Foveal Image Processing

6.1 Introduction

The unresolved target localization task was predominantly used in the preceding chapters to simplify the analyses of foveal system performance and algorithms. This simplification resulted from the fact that object shape or size did not have to be considered (other than confining the object to being a bright minimum sized dot), and permitted closed analytical expressions predicting system behavior and performance. This chapter addresses the more general problem of processing resolved objects, i.e., objects extending over more than one pixel, resulting in foveal perceptions with shape and perhaps texture information. Tasks include object identification and characterization.

The existing "tools of the trade," such as filtering, edge detection, and region growing, are designed primarily for space invariant systems, and unfortunately do not lend themselves to the processing of non-uniformly sampled data. This chapter presents a novel hierarchical approach to the processing of foveal data which supports much of the wide repertoire of image processing tools while retaining the information selective and data reduction features of foveal vision.

The resolution required to analyze a scene feature varies depending on the feature (e.g., its scale and bandwidth) and the task (e.g., classification by vertex, or by silhouette analysis). An important function of the hierarchical approach to foveal data processing is the estimation of the resolution required to process a detected scene feature, and the "request" for additional (higher bandwidth) information on specific regions of the scene when necessary. A new technique for the implementation of gaze control strategies is presented in this chapter which services these requests, and can satisfy multiple requests simultaneously with one foveation.

6.2 Hierarchical Foveal Data Representation

As discussed in Section 4.9, there are three dimensions implicitly associated with a rexel value (*x* location, *y* location, and acuity), whereas only two dimensions are associated with a pixel value (*x* and *y* location). Image arrays, being two dimensional data structures, properly represent the implicit pixel data, e.g., the sample value of the scene adjacent to another sample is the neighboring pixel, or array element. However, rexel data cannot be stored as compactly in a uniform two dimensional data structure. Rexel data could be represented on a non-uniform two dimensional data structure such as a 2-D linked list, but location would have to be explicitly stored, and the list would have to be restructured upon each integration of new frame data. Furthermore, the resulting structure remains space variant and incompatible with most image processing tools. This approach is nevertheless possible, and may be a candidate application of neural networks. Supporting this is neural modeling of the front-end processing in vertebrate vision [Curla83], [Kunik83], [Oguzt83].

An alternative to foveal data representation is the use of a three dimensional data structure in which the three-tuple of implicit information accompanying each rexel value is represented by the value's location within the structure (i.e., one data structure dimension for each dimension of implicit information). Rexel values can be stored at a location in the structure corresponding to the planar location of the rexel within the field-of-view (first two coordinates) and the rexels acuity (third coordinate).

6.2.1 Image Pyramids

Such three dimensional data structures have been the topic of considerable research in the field of image processing, although not in the context of processing space variant raw data. The structures are collectively called image pyramids. One motivation behind pyramids is to overcome the problems of feature scaling in image analysis. An *image pyramid data structure*, or *pyramid*, is a sequence of representations (image frames) generated from a uniformly sampled image at progressively lower resolutions (Figure

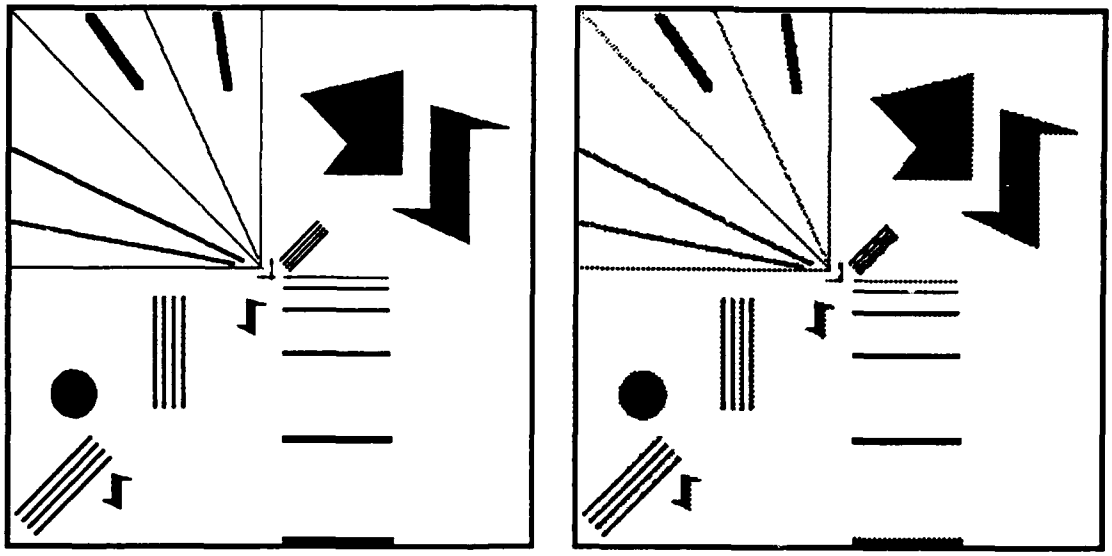
6.2.1-1). The size of the representation is reduced commensurately with resolution. Each frame forms a level within the pyramid, and the input sensor frame forms the base level (Figure 6.2.1-2). Note that each level of the pyramid is a conventional image frame at uniform resolution. The levels of Figure 6.2.1-2 are scaled to a uniform cell size. An alternate representation of the image pyramid data structure is to scale each level to their field-of-view. The resulting representation is a rectangular parallelepiped where each cell therein is "stretched" to cover the field-of-view from which its value is computed (Figure 6.2.1-3).

Image processing tools are applied at lower levels of resolution (higher pyramid levels) so as to exploit the computational savings associated with operating on smaller frames. Computational savings are obtained by searching for features in lower resolution frames, and testing localized hypotheses in spatial subsets of the higher resolution frames, thus avoiding searching throughout the entire base frame of the pyramid. Also, the overhead of wide ranging feature scaling is alleviated by working in the appropriate level of the pyramid which "normalizes" feature sizes.

The pyramid is typically generated in a bottom-up fashion from the original image, which forms the base of the pyramid. Different types of pyramids have been developed, varying in their bottom-up generation algorithm and resulting data structure. Two principle types of pyramid data structures are *Gaussian* and *Laplacian* pyramids. Gaussian pyramids are formed by assigning the value of a cell above the base (parent cell) the average of its direct descendent cells at the lower level (sibling cells) weighted by a fixed kernel centered at the parent cell. The generalized Gaussian pyramid bottom-up algorithm is

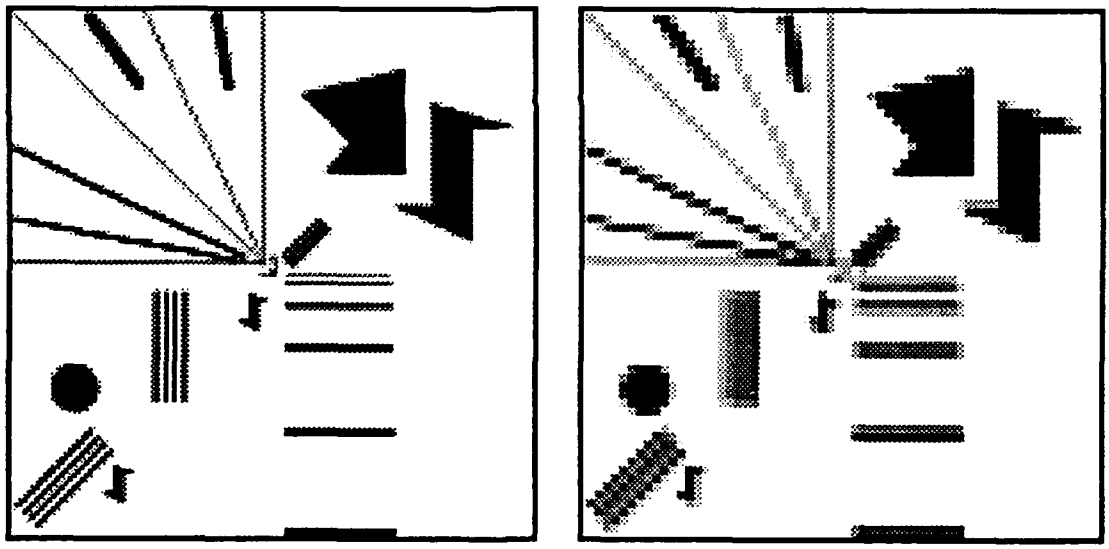
$$ {}^{k}G_{i,j} = \sum_{n_1=0}^{a-1} \sum_{n_2=0}^{a-1} {}^{k-1}w_{n_1,n_2} \, {}^{k-1}G_{mi+n_1,mj+n_2} \qquad (6\text{-}1) $$

where k is the level index (k=0 represents the base level), i and j are the indices of the cell at level k (i,j=0 being the corner cell, $i,j \geq 0$), ${}^{k}G_{ij}$ is the value of the pyramid at planar location i,j of level k, ${}^{k}w_{ij}$ is the value of the kernel of size $a \times a$ at location i,j, and m is the order of level reduction (i.e., if the base level is of size $N \times N$, the size of level k is $\frac{N}{m^k} \times \frac{N}{m^k}$). If $a>m$, then the kernels of adjacent parent cells overlap.



a. Level 0 (original image, 512×512 pixels)

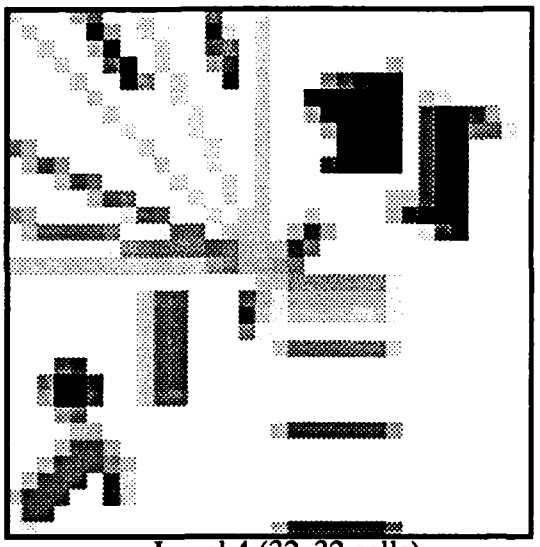
b. Level 1 (256×256 cells)



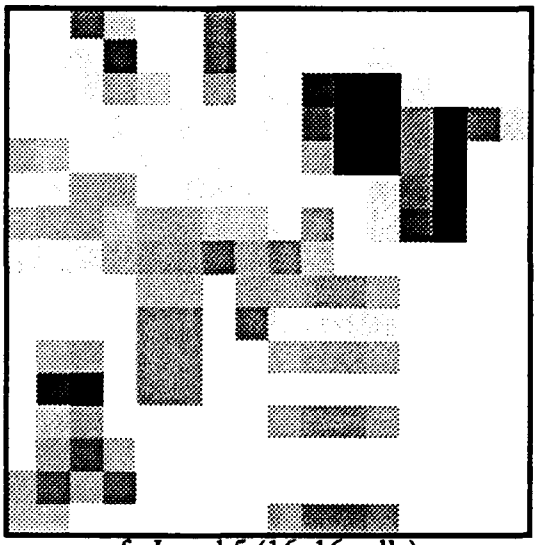
c. Level 2 (128×128 cells)

d. Level 3 (64×64 cells)

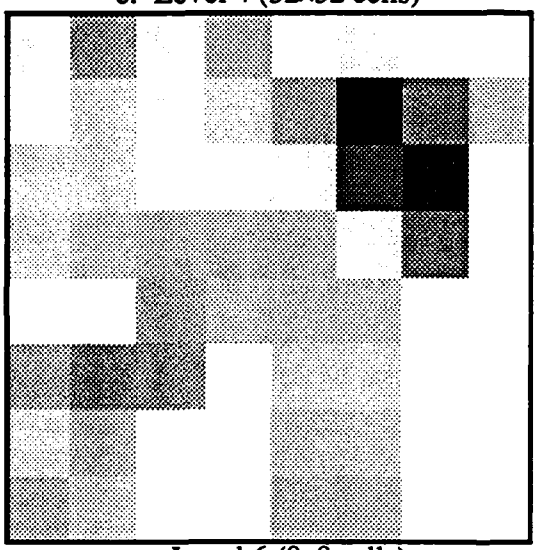
Figure 6.2.1-1. Levels of an image pyramid. Example uses the 512×512 pixel image shown in (a). (continued on next page)



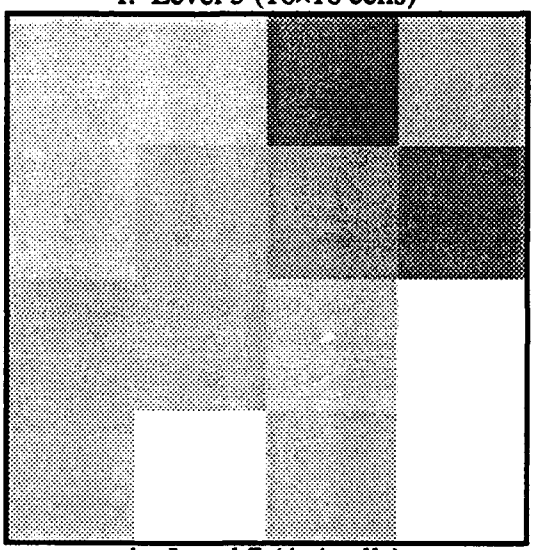
e. Level 4 (32×32 cells)



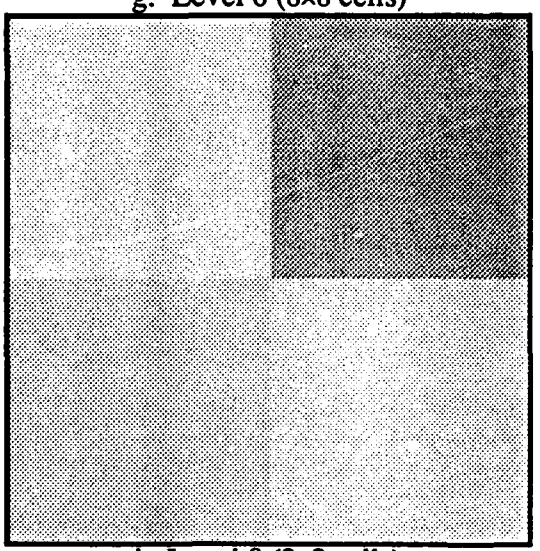
f. Level 5 (16×16 cells)



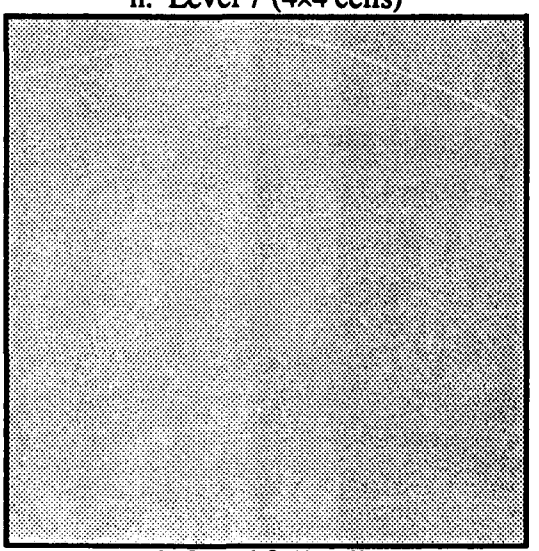
g. Level 6 (8×8 cells)



h. Level 7 (4×4 cells)



i. Level 8 (2×2 cells)



j. Level 9 (1 cell)

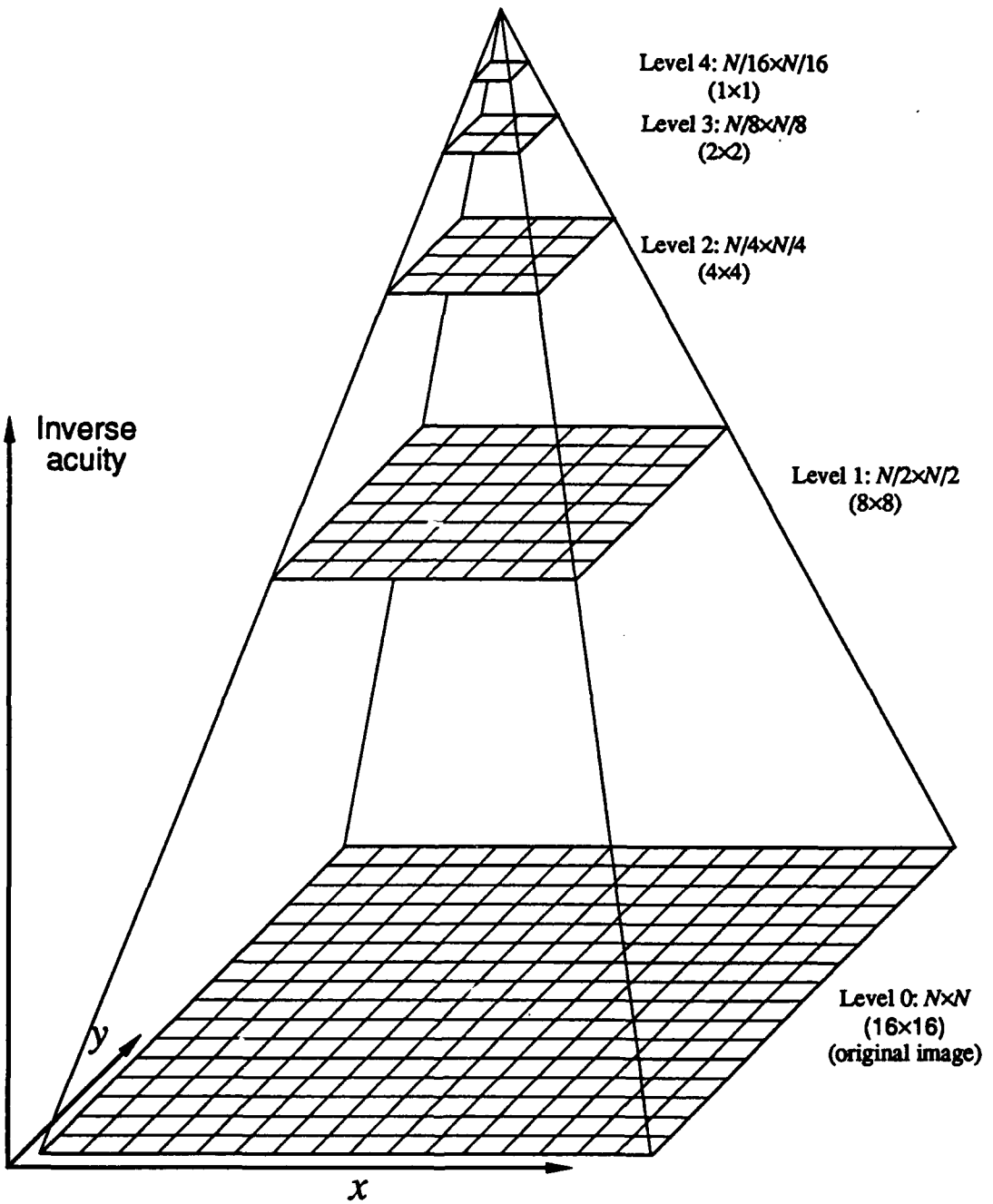


Figure 6.2.1-2. Conventional representation of image pyramid data structure over a 16×16 pixel base.

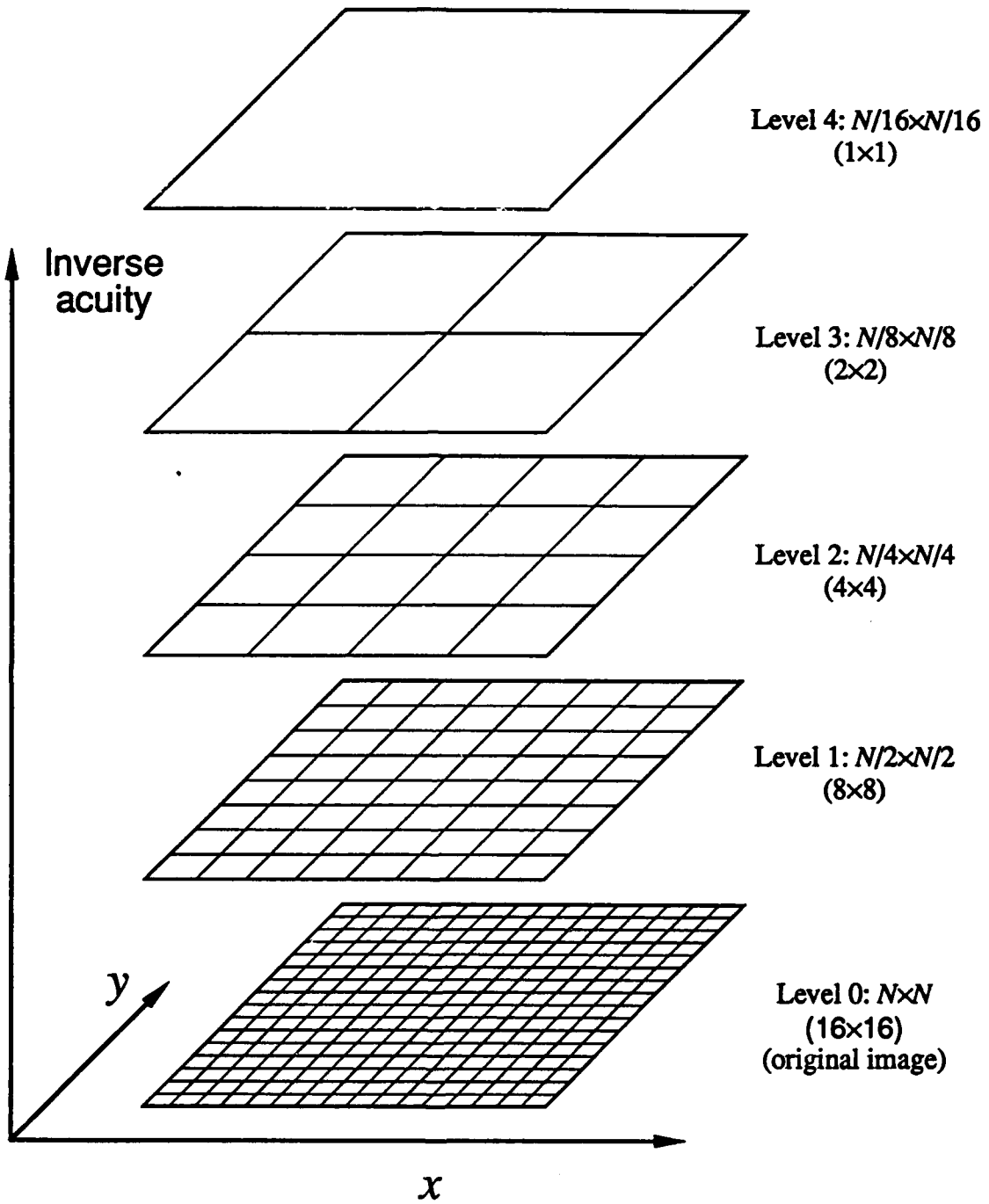


Figure 6.2.1-3. Alternate representation of image pyramid data structure over a 16×16 pixel base.

By using different kernels, (e.g., linear lowpass filters, nonlinear filters such as median or maximum value operators), the generated levels can emphasize different attributes of the base image so as to better support the particular vision task. Typically, the kernel of the Gaussian pyramid is a lowpass filtering operator commensurate with the reduction in level size. The pyramids store more data than the original image, but the increase is small, and can be more than compensated by the computational savings from manipulating less data.

The most common Gaussian pyramid is characterized by a=m=2, and is formed by averaging a 2×2 group of sibling cells into a parent cell:

$$ {}^{k}G_{i,j} = 0.25\left({}^{k-1}G_{2i,2j} + {}^{k-1}G_{2i+1,2j} + {}^{k-1}G_{2i,2j+1} + {}^{k-1}G_{2i+1,2j+1}\right) \tag{6-2} $$

A given pyramid level has half the linear dimensions (one fourth the data) of the level underneath. The linear dimension of the base $N \times N$ is an integer power of two so as to obtain a complete pyramid ending with a single value at the top level representing the global average of the base image. The ranges of indices supported by this bottom-up algorithm, i.e., the three-tuples of implicit information, are

$$ k \in [0,\ldots,\log_2 N] \tag{6-3} $$

$$ i(k), j(k) \in [0,\ldots,2^{N-k}-1] \tag{6-4} $$

The position with respect to the base coordinates of the center of a cell at level k is given by

$$ c_i(k) = 2^k(i+0.5) \tag{6-5} $$

$$ c_j(k) = 2^k(j+0.5) \tag{6-6} $$

Such a Gaussian pyramid was used to generate Figure 6.2.1-1. The acuity at level k is 2^k times less than the maximum system acuity, which is that of the base image. The total amount of data in the pyramid is

$$ D_{total} = \sum_{k=0}^{\log_2 N} 2^k \times 2^k = \frac{4^{1+\log_2 N}-1}{3} = \frac{4}{3}N^2 - \frac{1}{3} \tag{6-7} $$

which is approximately 33% greater than the original (base) image.

The levels of the Laplacian pyramid are reduced versions of the base image filtered by a bandpass filter (also called a difference of Gaussians, sombrero, or Laplacian filter), with a progressively lower bandpass center frequency used at higher pyramid levels [Burt83]. Figure 6.2.1-4 illustrates the Laplacian pyramid corresponding to Figure 6.2.1-1.

The Laplacian pyramid data structure may be formed from a Gaussian pyramid of the same dimensions. Specifically, the level of a Laplacian pyramid may be formed by taking the difference between the corresponding level of a Gaussian pyramid and the expanded version of the next higher level (thus both array arguments to the differencing operation are of the same size). Expansion is performed by replicating each cell into $m \times m$ cells of the same value, i.e.,

$$\overset{k}{\bullet}G = EXPAND\left({}^{k}G\right) \tag{6-8}$$

where

$$\overset{k}{\bullet}G_{i,j} = {}^{k}G_{\left\lfloor \frac{i}{m} \right\rfloor, INT\left\lfloor \frac{j}{m} \right\rfloor} \tag{6-9}$$

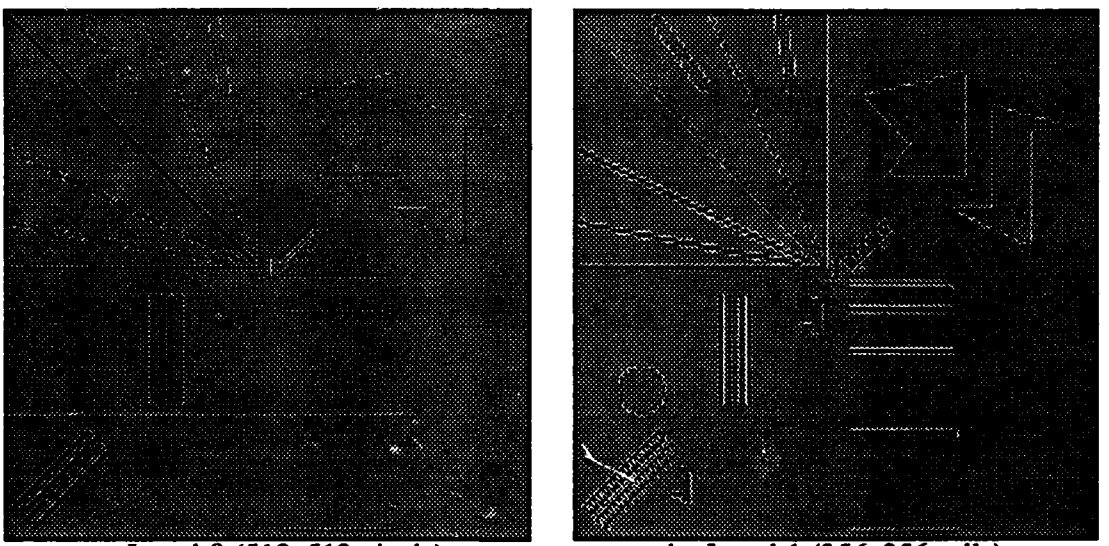
This approach to the generation of the Laplacian pyramid results in a data structure which is one level shorter than the Gaussian pyramid employed in the generation process. More accurate (and computationally intensive) versions of Laplacian pyramids may be formed by using interpolation rather than replication [Burt83].

6.2.2 Storing Foveal Frames in Pyramids: The Foveal Manifold

When the original image is obtained through foveal sampling, the conventional approach to pyramid generation does not apply. The rexels values cannot be represented in a two dimensional base array from which the pyramid data structure can be generated, nor do all the rexels correspond to the same pixel resolution of the base. However, if the acuity and location of the rexels correspond to acuity and location indices supported by the pyramid, then the individual rexel values can be stored in the pyramid. The values will not all be at the base but throughout a subset of the pyramid data structure. This requires that

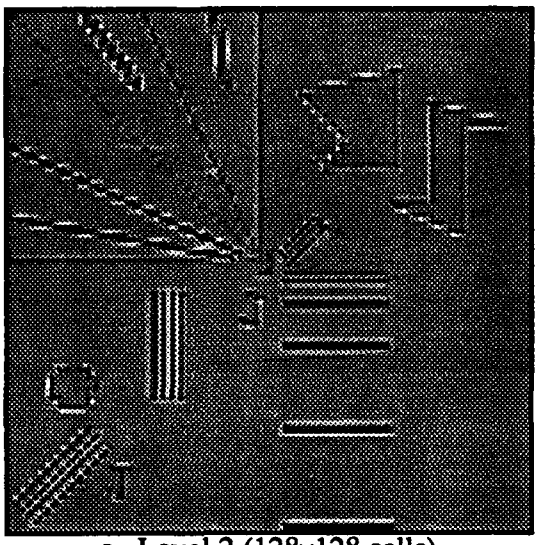
the bottom-up pyramid generation process match in a certain sense the rexel size function of the foveal sensor geometry. Specifically, the generation process of pyramid cells from frame pixels must match the generation process of foveal rexels from scene pixels.

For example, consider the Gaussian pyramid constructed using (6-2) and a frame from a foveal sensor with undivided exponential geometry. The acuity values taken on by the pyramid levels coincide with those taken by the different rexel sizes of the foveal geometry, namely integer powers of two reductions from the maximum system resolution (corresponding to the fovea, or central 4×4 rexels, of the exponential geometry, and the base level of the pyramid). As in previous analyses, maximum resolution shall be normalized to that of a single pixel (1×1 rexel). Furthermore, rexel locations within the exponential geometry are implicitly represented by the locations of values within the pyramid. The central 4×4 locations of the pyramid base can store the four central rexels (first ring) of the exponential pattern, the second pyramid level can store the second ring of rexels, and so forth. The storage in the Gaussian pyramid of an exponential foveal frame with r rings allocates values to the first r levels. Figure 6.2.2-1 illustrates the allocation of rexel values within the Gaussian pyramid.

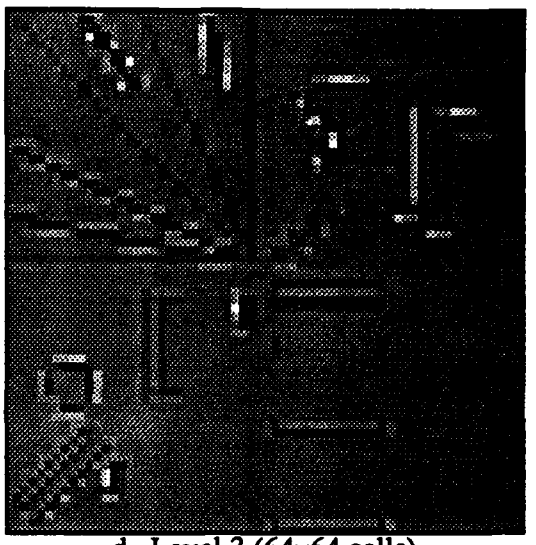


a. Level 0 (512×512 pixels) b. Level 1 (256×256 cells)

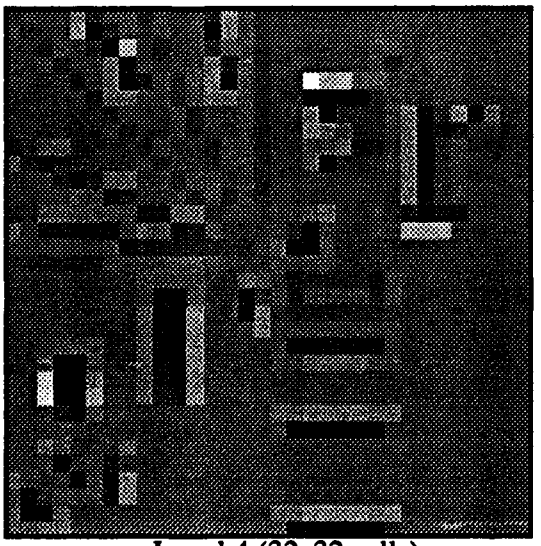
Figure 6.2.1-4. Levels of a Laplacian pyramid. Levels are computed from the Gaussian pyramid of Figure 6.2.1-1. (continued on next pages)



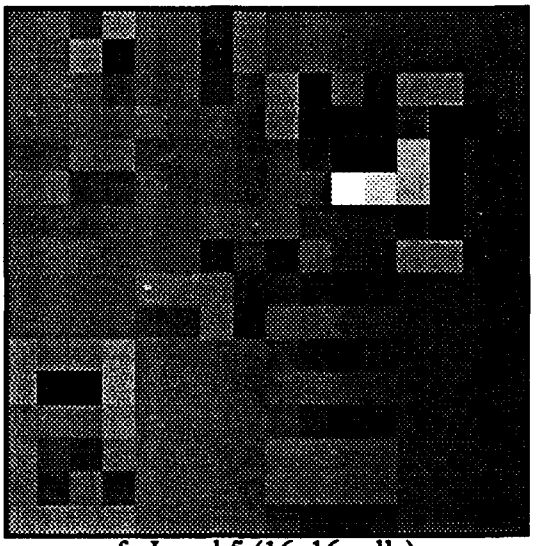
c. Level 2 (128×128 cells)



d. Level 3 (64×64 cells)

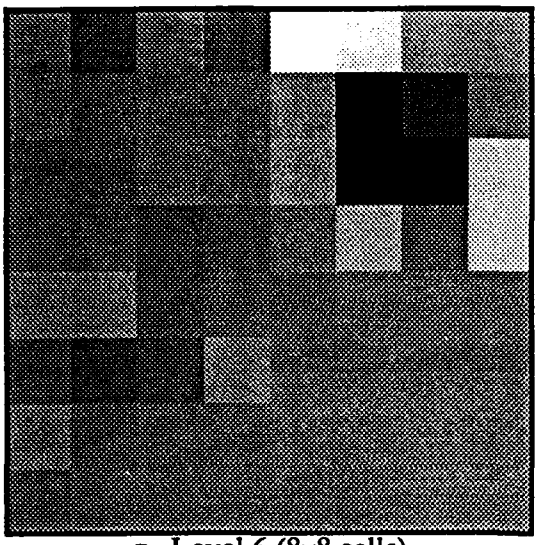


e. Level 4 (32×32 cells)

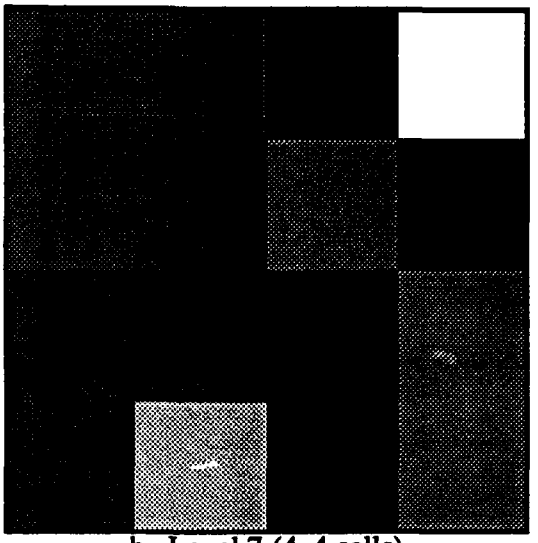


f. Level 5 (16×16 cells)

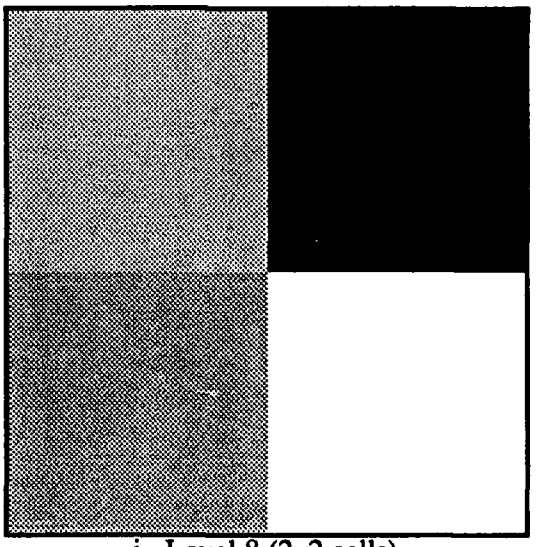
Figure 6.2.1-4. Levels of a Laplacian pyramid (continued from previous page).



g. Level 6 (8×8 cells)



h. Level 7 (4×4 cells)



i. Level 8 (2×2 cells)

Figure 6.2.1-4. Levels of a Laplacian pyramid (continued from previous page).

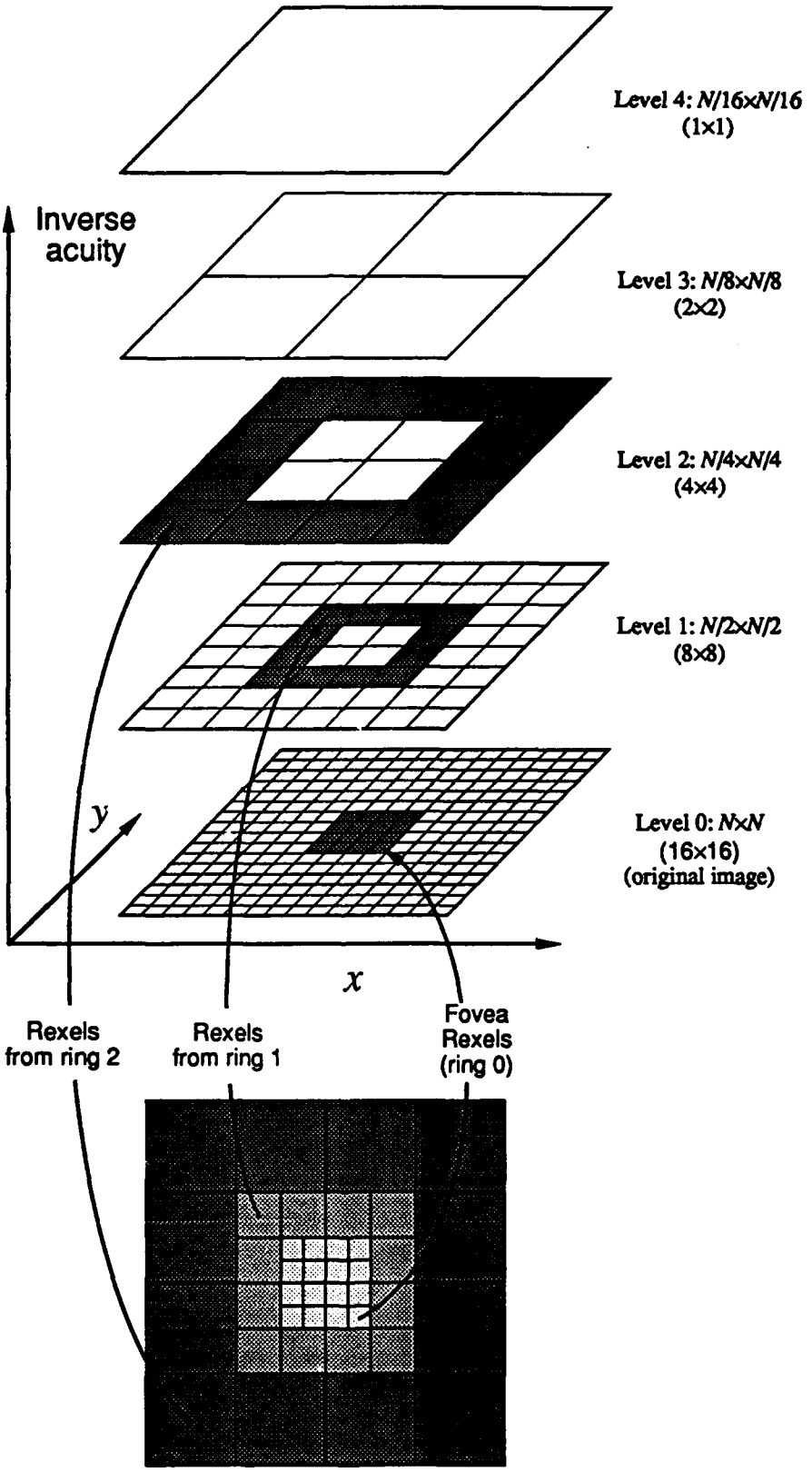


Figure 6.2.2-1. Mapping of exponential geometry rexels into the image pyramid. Wider fields-of-view (pyramid bases) support additional sensor rings.

The mapping of rexels to the image pyramid is possible only when the rexels sizes are integer powers of 2 and rexel boundary lines are maximally preserved. These attributes are unique to the exponential pattern and its subdivisions. Most rings in the linear pattern have rexel sizes not supported by the pyramid. Uniformly subdivided exponential patterns are supported by the pyramid because rexel dimensions are still related by a power of two. Non-uniform subdivisions are supported as long as the factors are also related by a power of two, as formulated in Section 3.5.1. Otherwise, not all the resulting rexel sizes will be supported by the pyramid. In general, rexel values and cell values match exactly whenever the rexel impulse response matches the pyramid kernel (overlapping receptive fields are consistent with overlapping kernel pyramids).

The three dimensional representation of the pyramid with the exponential rexel frame illustrates a locus of rexel values which forms the shell of an upside-down pyramid. This pyramid shell, named the *foveal manifold*, contains all the rexel values (Figure 6.2.2-2). The height of the manifold is determined by the number of major rings of the sensor. From (3-13) it is seen that the first r rings of the undivided exponential pattern cover perfectly an area of $2^{r+1} \times 2^{r+1}$. Since the linear dimensions $N \times N$ of the field-of-view are some integer power of two, then an integer number of rings will fit exactly in the pyramid. Specifically, if the field-of-view is $2^R \times 2^R$, then R-1 rings will fit in the pyramid, and the foveal manifold is R-1 levels high (registering in levels k=0 to k=R-2) within a pyramid of R+1 levels (k=0 to R). Each level of the foveal manifold contains a square ring of 12 rexel values except for the bottom level which contains the 4×4 foveal rexels.

Since rexel subdivision does not affect major ring boundaries, an integer number of rings of a subdivided exponential pattern likewise fits inside the pyramid. In this case, the first level of the resulting foveal manifold contains $4d \times 4d$ values, where d is the uniform subdivision factor, and all other manifold levels contain $12d^2$ rexel values arranged as d concentric square rings (Figure 6.2.2-3). The data distribution of the manifold is obtained just as the data for the sensor geometry: the rexels of the undivided manifold/geometry are subdivided into $d \times d$ elements, which are then scaled upward by a factor d to retain the same maximum resolution. The number of major rings m' (groups of d rexel rings with uniform rexel size) in a subdivided exponential geometry covering an overall area of $2^R \times 2^R$ is the same as the number of rexel rings m in an undivided exponential geometry covering an area of $d^{-1}2^R \times d^{-1}2^R$, and the foveal manifold is R-1-$\log_2 d$ levels high (registering in levels k=0 to k=R-2-$\log_2 d$):

$$m' = \frac{\log_2\left(\frac{2^{2R}}{d^2}\right)}{2} - 1 = R - \log_2 d - 1 \qquad (6\text{-}10)$$

or, if the subdivision factor is an integer power of two $d=2^{\sigma}$,

$$m' = R - \sigma - 1 \qquad (6\text{-}11)$$

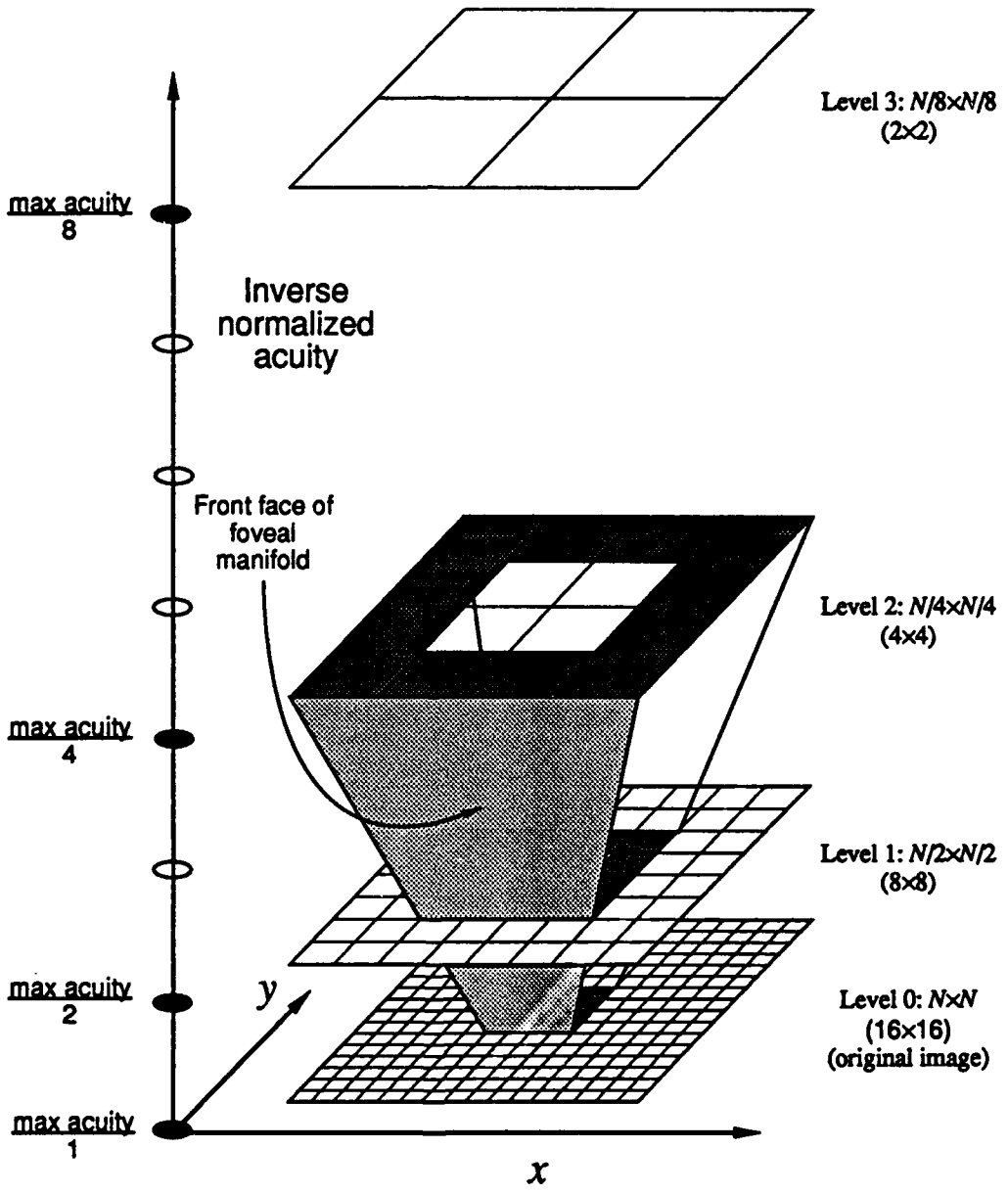


Figure 6.2.2-2. The foveal manifold of an undivided exponential sensor frame.

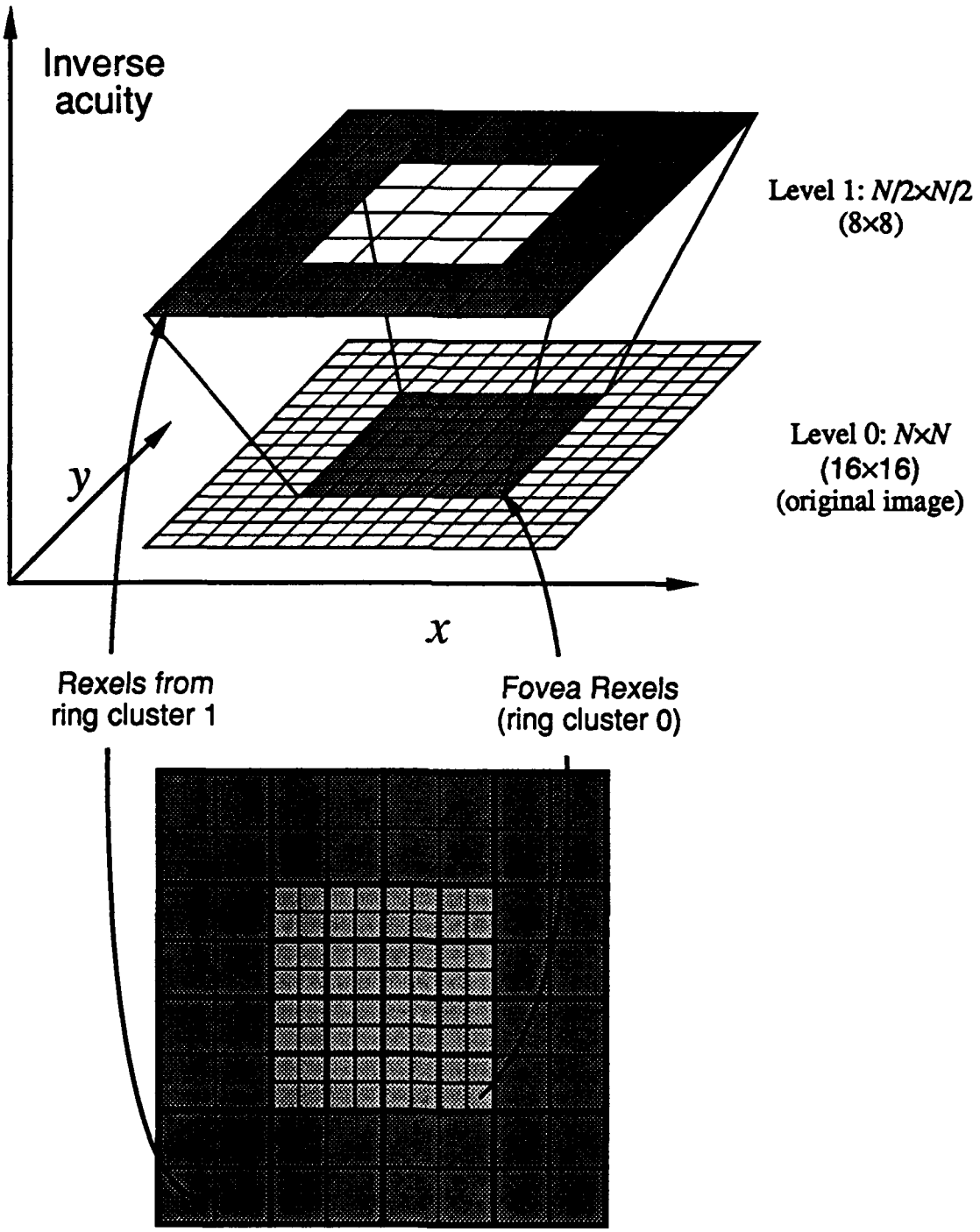


Figure 6.2.2-3. Foveal manifold of a subdivided (factor of 2) exponential frame.

6.2.3 The Foveal Polygon Hierarchical Data Structure

Conventional pyramid vision algorithms require the entire data structure of the pyramid to be computed in a bottom-up fashion from the uniform resolution input image (the pyramid base) before further analysis can be performed. A similar data structure synthesis operation can be performed on rexel data. A principle difference between data structure generation from pixels versus from rexels is that the former commences with a complete base level, while the latter commences with the multilevel foveal manifold.

The data structure generated from the rexel data is a subset of an image pyramid because the manifold does not complete any pyramid level, and because the bottom-up generation process can only compute those cell values above the foveal manifold (Figure 6.2.2-2). This is to be expected because only a frame of data at maximum resolution across the entire field-of-view can support the generation of the entire pyramid. The data structure generated from the rexel data is called the *foveal polygon.*

The foveal polygon resulting from the application of the Gaussian pyramid generation technique defined by (6-2) on the foveal manifold is called the *Gaussian foveal polygon.* The values of the center 2×2 cells of level k=1 framed by the manifold rexel values are computed by averaging the corresponding sibling cells from level k=0. No other cells in level k=1 can be generated. Thus, after bottom-up generation from the data of an undivided exponential geometry frame, level k=1 contains 12 measured values (foveal manifold cells) and 4 computed values.

The data at level k=1 of the Gaussian polygon is similar to that of level k=0 in that the central 4×4 cells are defined and there is no foveal data to support any more cells at those level. The levels differ in that at k=1, 75% of the values are observed and 25% derived. The foveal sensor can be considered as a realization of (6-2) implemented at the sensor, producing data values which equal the result of the sibling cell averaging operation, instead of the sibling cells themselves.

The generation of level k=2 follows the same approach as k=1. The values of the 2×2 cells of level k=2 framed by the manifold rexel values are obtained by averaging the corresponding sibling cells from level k=1. The result, again, is a level with the central 4×4 cells initialized. Even though the number of cells supported by foveal data at each level is

constant, the relative completion of each level increases with k because the size (in cells) of the hierarchical levels become smaller.

This bottom-up generation process continues up the hierarchical levels. At level k=R-2, all the cells at this level are assigned values. This is the highest level in the Gaussian polygon containing rexel data (the foveal manifold ends at this level), and it has the acuity (and rexel values) of the outermost ring of the undivided exponential geometry. This level is called the *waist* of the foveal polygon. The polygon levels at and above the waist are identical to the corresponding levels of a Gaussian pyramid formed from a corresponding uniresolution frame because the waist covers the entire field-of-view.

When the exponential foveal pattern is uniformly subdivided by a factor d, the manifold widens as levels are assigned $12d^2$ rexel values instead of only 12, and the base is assigned $16d^2$ values. The top of the foveal manifold (the polygon waist) is at level m', which from (6-10) is $\log_2 d$ levels lower than if the pattern were not subdivided (d=1). For each level with rexels (excluding the base), the Gaussian polygon generation process computes $4d^2$ cell values, for a total of $16d^2$ cell values per level. A necessary condition for a parent cell at or below the waist to have all four sibling cells underneath is that d must be an integer power of two.

The Gaussian foveal polygon data structure resulting from a subdivided exponential foveal sensor frame can be segmented into two regions (Figure 6.2.3-1):

1. The foveal polygon within the manifold, from level k=0 to level k=R-$\log_2 d$-2, containing all the rexel measurements and bottom-up computed values at a 3:1 ratio.

2. Conventional pyramid section, consisting of levels k=R-$\log_2 d$-2 to k=R.

A third region can be defined as that within a corresponding pyramid but outside the foveal polygon. The size D_o (in cells) of this third region is computed by subtracting from the size of each pyramid level from k=0 to k=R-$\log_2 d$-2 the size of the polygon at that level, and summing the terms:

$$D_o = \sum_{i=0}^{R-\log_2 d-2}\left[\left(\frac{2^R}{2^i}\right)^2 - 16d^2\right]$$

$$= (2^R)^2 \sum_{i=0}^{R-\log_2 d-2}\left[4^{-i}\right] - 16d^2(R-\log_2 d-1)$$

$$= \frac{4}{3}4^R - 16d^2\left(R-\log_2 d-\frac{2}{3}\right) \qquad (6\text{-}12)$$

As expected, D_o decreases as d is increased because the foveal sensor frame is more "uniform", and the polygon flares wider (by a factor d^2) to cover more volume in less levels. Indeed, if $d=2^{R-2}$, then the subdivided fovea covers the entire field-of-view and $D_o=0$.

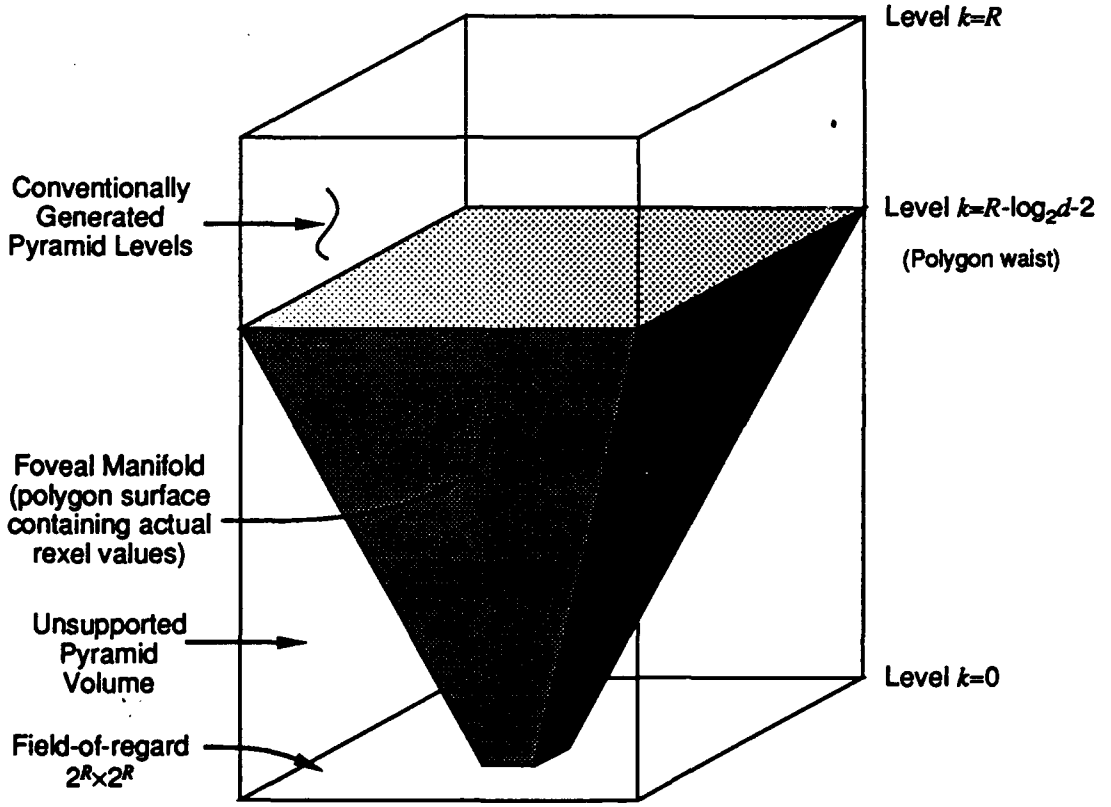


Figure 6.2.3-1. Components of the foveal polygon.

Comparing the size of the foveal polygon, including the number of cells within the manifold D_f and the conventional pyramid section D_a above, to the number of rexels in the foveal sensor frame A_r indicates the cost of using the pyramid in terms of additional data generated. The number of cells in each level of the polygon within the manifold is $16d^2$. Given r major rings,

$$D_f = 16rd^2 \tag{6-13}$$

The size of the polygon above the waist is that of a conventional pyramid with base $2d \times 2d$ (6-7):

$$D_a = \frac{4}{3}4d^2 - \frac{1}{3} \cong \frac{16}{3}d^2 \tag{6-14}$$

The total size of the polygon data structure is

$$D_f + D_a = 16rd^2 + \frac{4}{3}4d^2 - \frac{1}{3} \cong 16d^2\left(r + \frac{1}{3}\right) \tag{6-15}$$

Thus, the ratio of total polygon size to that of the sensor frame is

$$\frac{D_f + D_a}{A_r} = \frac{16d^2\left(r + \frac{1}{3}\right)}{d^2(4 + 12r)} = \frac{4r + \frac{4}{3}}{3r + 1} = \frac{4}{3} \tag{6-16}$$

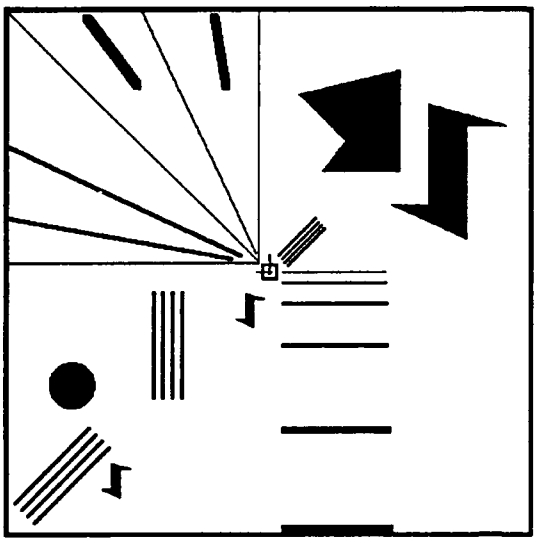
which is the same ratio between Gaussian pyramid and uniform sensor frame size. Consequently, the savings in foveal frame data with respect to uniform sensor frames, derived in Chapter 3, also apply to the resulting hierarchical data structure. The fact that this amount of data is manipulated in the bottom-up generation process does not imply that it is all processed in the ensuing, more computationaly intensive feature analyses. Indeed, here is where the benefits of hierarchical image processing are reaped: at low resolution, the analyses algorithms can define regions of interrogation, and then conduct the analyses at the corresponding regions in higher resolution levels, so as to process a total amount of data which is less than that of the base level alone. The bottom-up generation process can be performed at an exhaustively concurrent fashion (processing distributed at the datum level), representing minimum overhead.

When the foveal system foveates to a relevant object in the scene, the proportion of the polygon data used in feature analyses can be greater than that of the pyramid data. This is because the data structures are generated from sensor frames which themselves can have different ratios of relevant information to total data size, as demonstrated in Chapters 3 and 4.

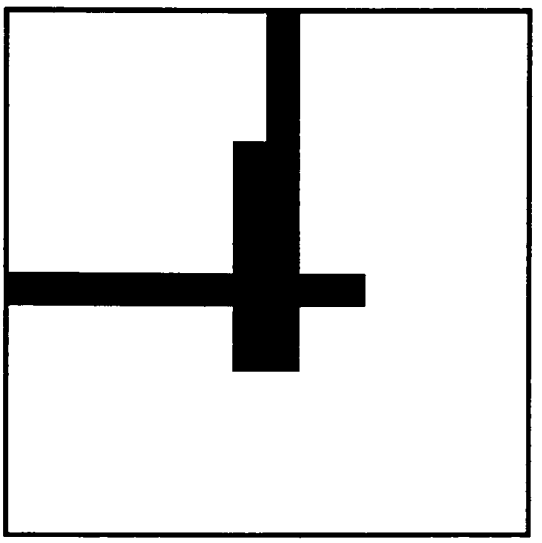
Figure 6.2.3-2 illustrates the first five levels (k=0 to k=4) of the Gaussian polygon built from the foveal data of a sensor frame registering the 512×512 pixel scene shown in Figure 6.2.1-1a. The sensor geometry is an exponential pattern subdivided by a factor d=4, as illustrated in Figure 3.5-1. The polygon consists of 16×16 cells at each level from k=0 to k=5. Level k=5 of a Gaussian pyramid with a 512×512 base is itself 16×16 cells. Thus, levels k=5 to k=9 are the same as in Figure 6.2.1-1.

Next to each polygon level in Figure 6.2.3-2 is the corresponding level of a pyramid generated from a uniform resolution image (i.e., Figure 6.2.1-1). Within each conventional pyramid level is a square frame delineating the area covered by the image polygon at that level. The polygon coverage increases as the level increases, up to k=5, at which point the polygon covers the entire level.

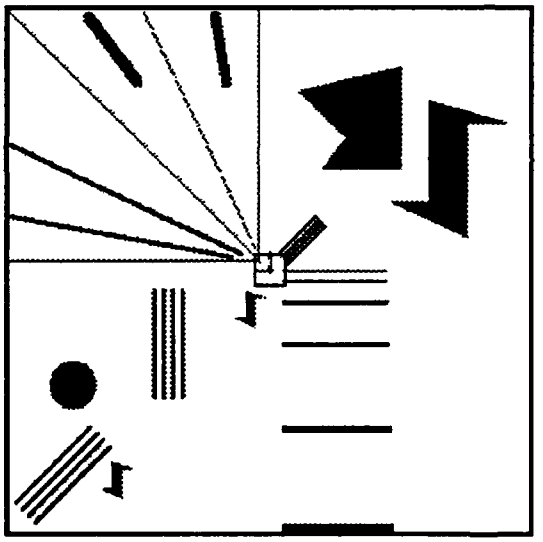
Just as the Gaussian foveal polygon is a subset of the Gaussian pyramid, a subset of the Laplacian pyramid can be derived from rexel data. This data structure is called the *Laplacian foveal polygon*, and can be formed by taking the difference between each Gaussian polygon level and the central $16d^2$ cells of the expanded version of the next polygon level above, a new data structure is Figure 6.2.3-3 illustrates the levels of the Laplacian polygon generated from the Gaussian polygon shown in Figure 6.2.3-2. The size and shape of the Laplacian polygon is approximately the same as that of the Gaussian polygon (as with the Laplacian pyramid, there is no 1×1 cell level at the top).



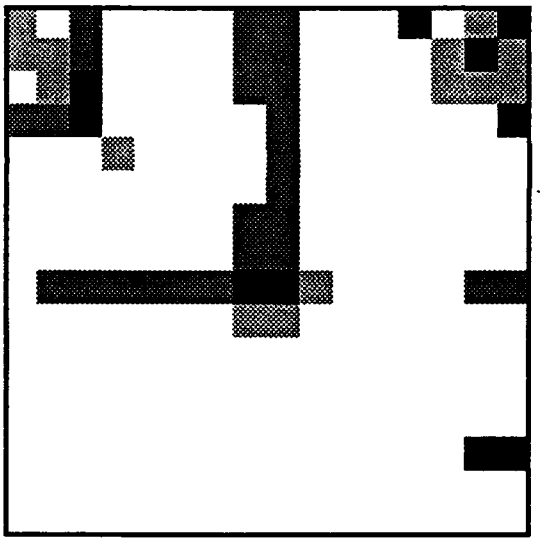
a. Pyramid level 0 (original image, 512×512 pixels)



b. Polygon level 0 (16×16 cells)

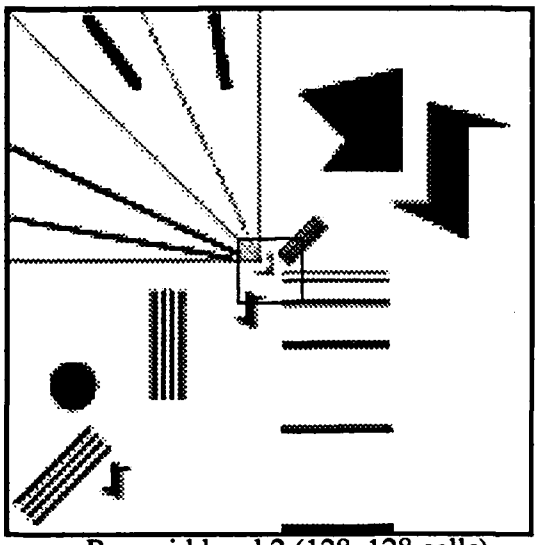


c. Pyramid level 1 (256×256 cells)

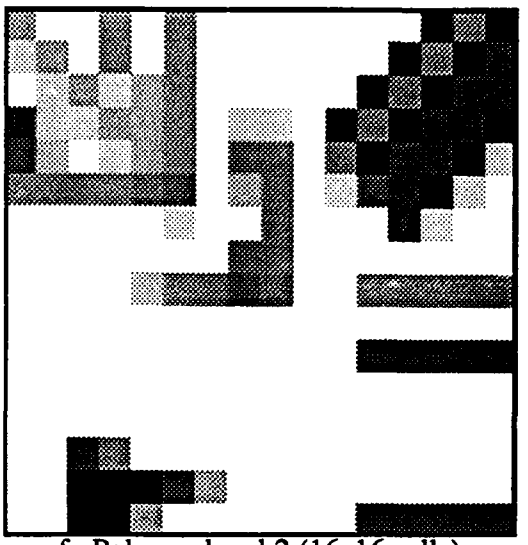


d. Polygon level 1 (16×16 cells)

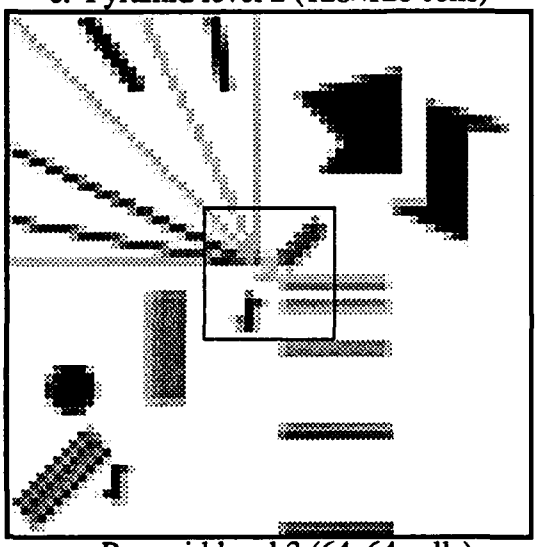
Figure 6.2.3-2. Levels of the Gaussian polygon. An exponential foveal pattern uniformly subdivided by a factor of 4 (4×4 rexels within each rexel of the original geometry) is used. To the left of each polygon level is the pyramid level formed from a uniform resolution pixel level sampling, denoting with a centered square frame the region supported by the polygon.



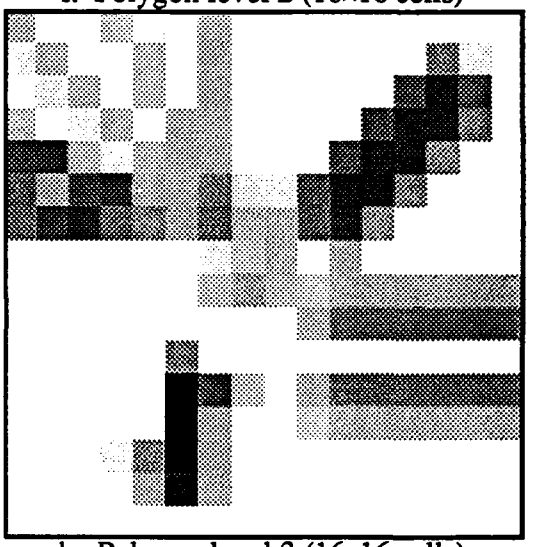
e. Pyramid level 2 (128×128 cells)



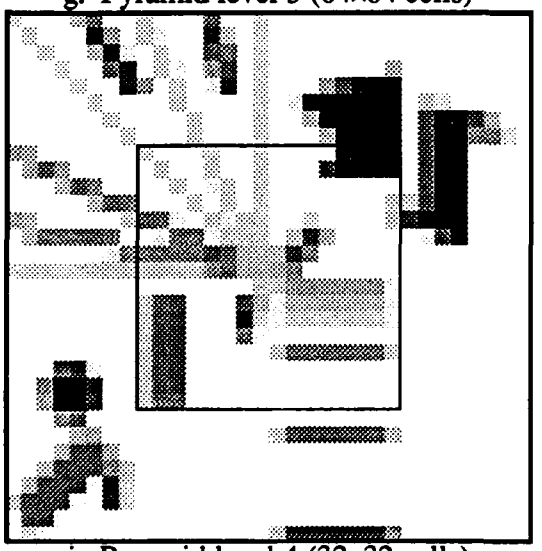
f. Polygon level 2 (16×16 cells)



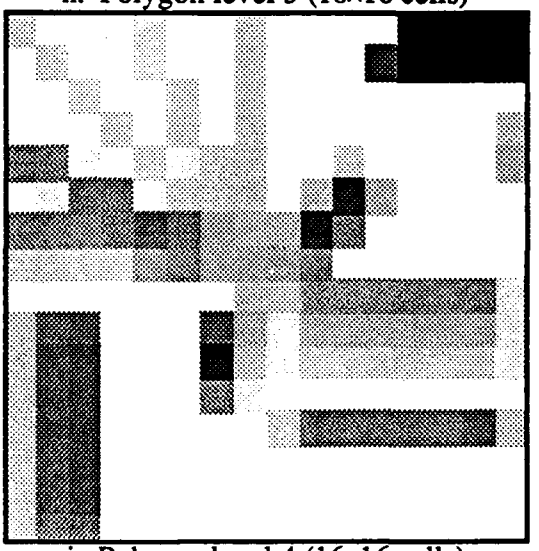
g. Pyramid level 3 (64×64 cells)



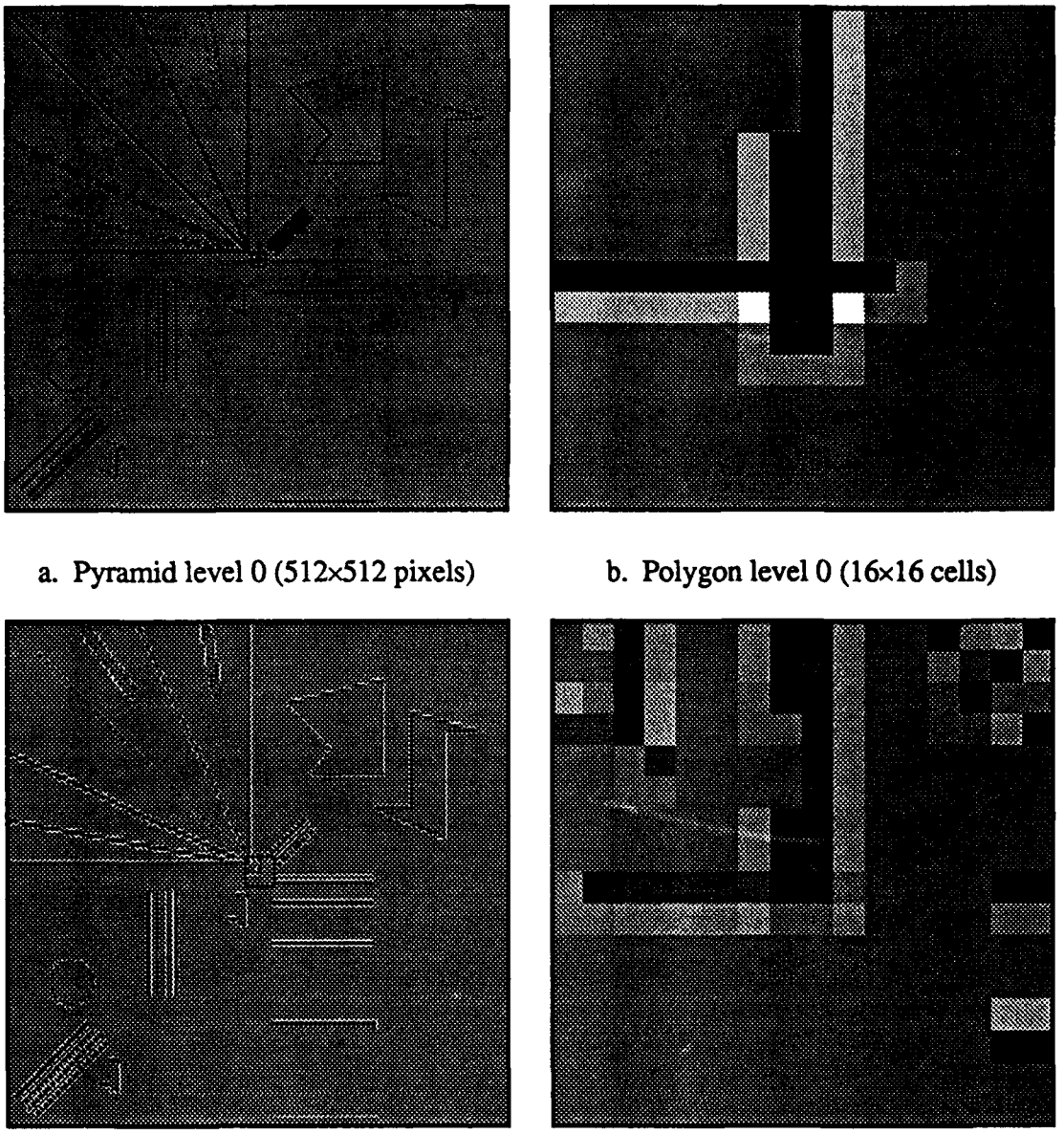
h. Polygon level 3 (16×16 cells)



i. Pyramid level 4 (32×32 cells)



j. Polygon level 4 (16×16 cells)



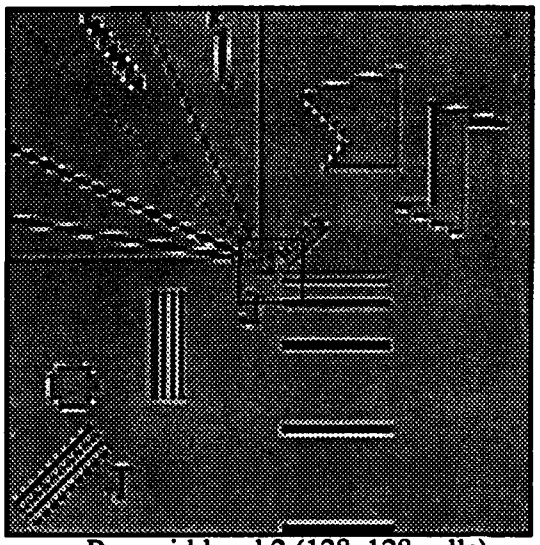
a. Pyramid level 0 (512×512 pixels)

b. Polygon level 0 (16×16 cells)

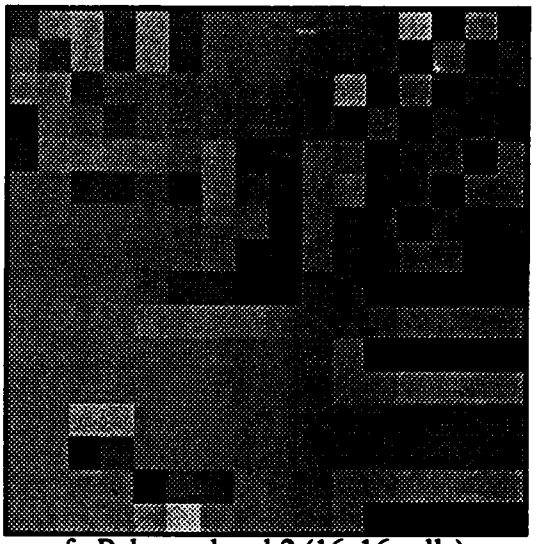
c. Pyramid level 1 (256×256 cells)

d. Polygon level 1 (16×16 cells)

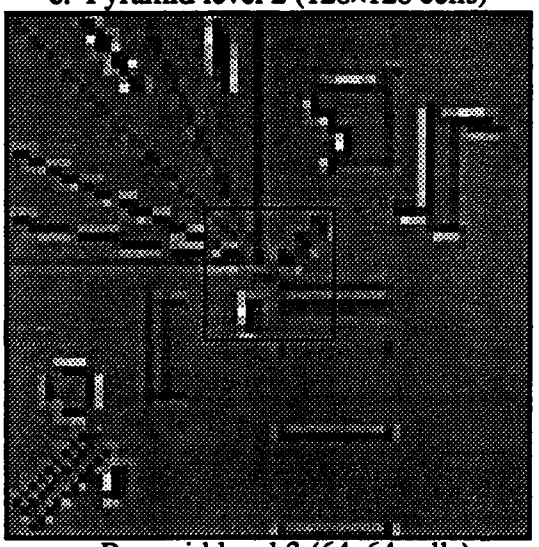
Figure 6.2.3-3. Levels of the Laplacian polygon. To the left of each polygon level is the Laplacian pyramid level formed from a complete Laplacian pyramid, denoting with a centered square frame the region supported by the polygon.



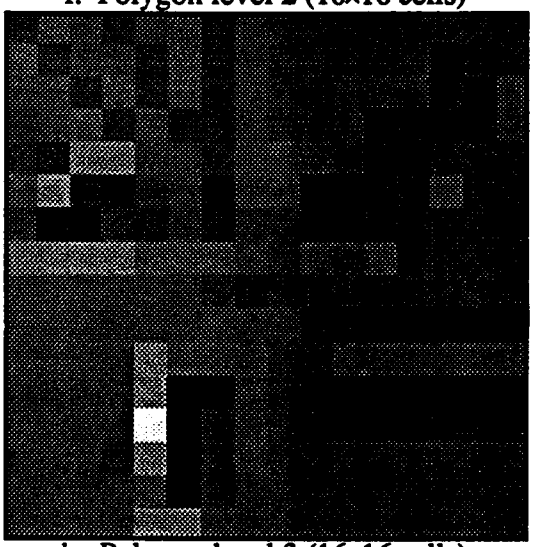
e. Pyramid level 2 (128×128 cells)



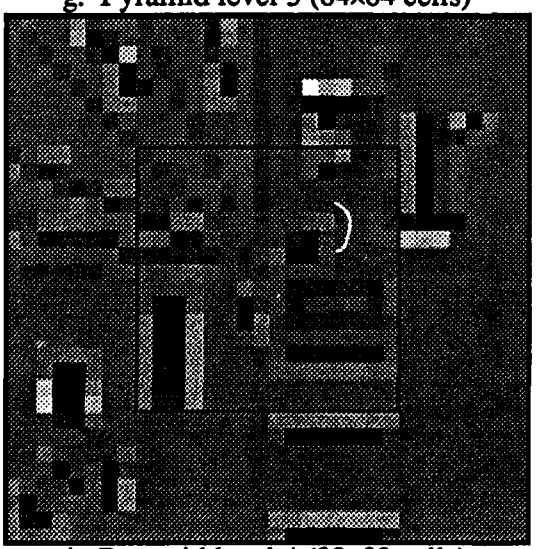
f. Polygon level 2 (16×16 cells)



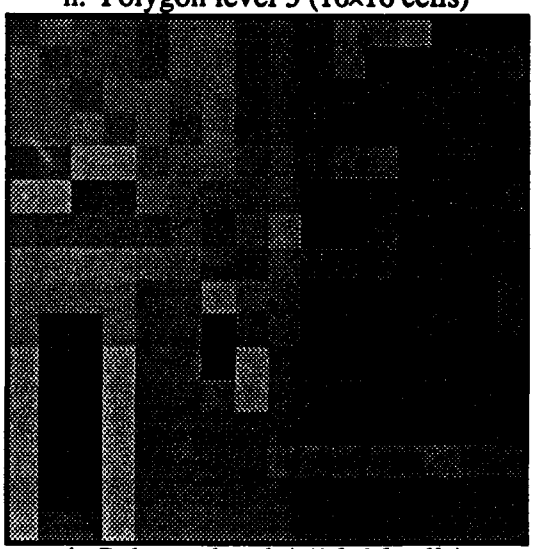
g. Pyramid level 3 (64×64 cells)



h. Polygon level 3 (16×16 cells)



i. Pyramid level 4 (32×32 cells)



j. Polygon level 4 (16×16 cells)

6.2.4 Foveal Polygon from Integrated Perceptions

In static scene applications, an integrated perception can be formed by fusing rexel values as discussed in Chapter 4. Data analysis is then performed on the integrated perception as opposed to the individual frames. A foveal polygon can be generated from the integrated perception, just as from an individual sensor frame. A key difference between the polygon from a sensor frame, discussed in the preceding section, and the polygon from an integrated perception is that the base of the former represents the field-of-view, while the base of the latter represents the field-of-regard.

The foveal polygon generated from an integrated perception after two foveations is illustrated in Figure 6.2.4-1. The rexel values of the two foveae are stored at the base level of the polygon. The remaining rexel data flares from the corresponding region of support at level k=0, resulting in two intersecting manifolds. Note that if the field-of-view were much smaller than the field-of-regard, the manifolds could possibly not intersect. However, foveal systems efficiently lend themselves to wide fields-of-view. The waist of an integrated perception polygon is the lowest level in which the entire field-of-regard is registered.

The Gaussian pyramid supports the implicit information accompanying a rexel from an exponential pattern centered over the pyramid base, which represents the system field-of-view. Such a foveation is a logical first registration. However, the coordinates for the centers of pyramid cells at levels above k=0 may not correspond to rexel centers if the optical axis is off-center. Figure 6.2.4-2 illustrates how by shifting the foveal axis by one pixel to the right, the spatial domain of the rexels (shaded regions) at levels above the base do not register with that of the pyramid cells.

This problem is analogous to that of pixel statistics correlation discussed in Section 4.3.1. Note how the rexels from the two different registrations illustrated in Figure 4.3.1-2 overlap perfectly in the sense that a rexel in one frame exactly overlaps a similarly sized rexel or a group of smaller rexels in the other frame. This condition, which minimizes pixel statistics correlation, is necessary for both frames to be represented simultaneously in a foveal polygon data structure without any interpolation of values (i.e., the direct mapping of rexel values to pyramid cells). Sufficiency is obtained when one of the frames is centered over the field-of-regard.

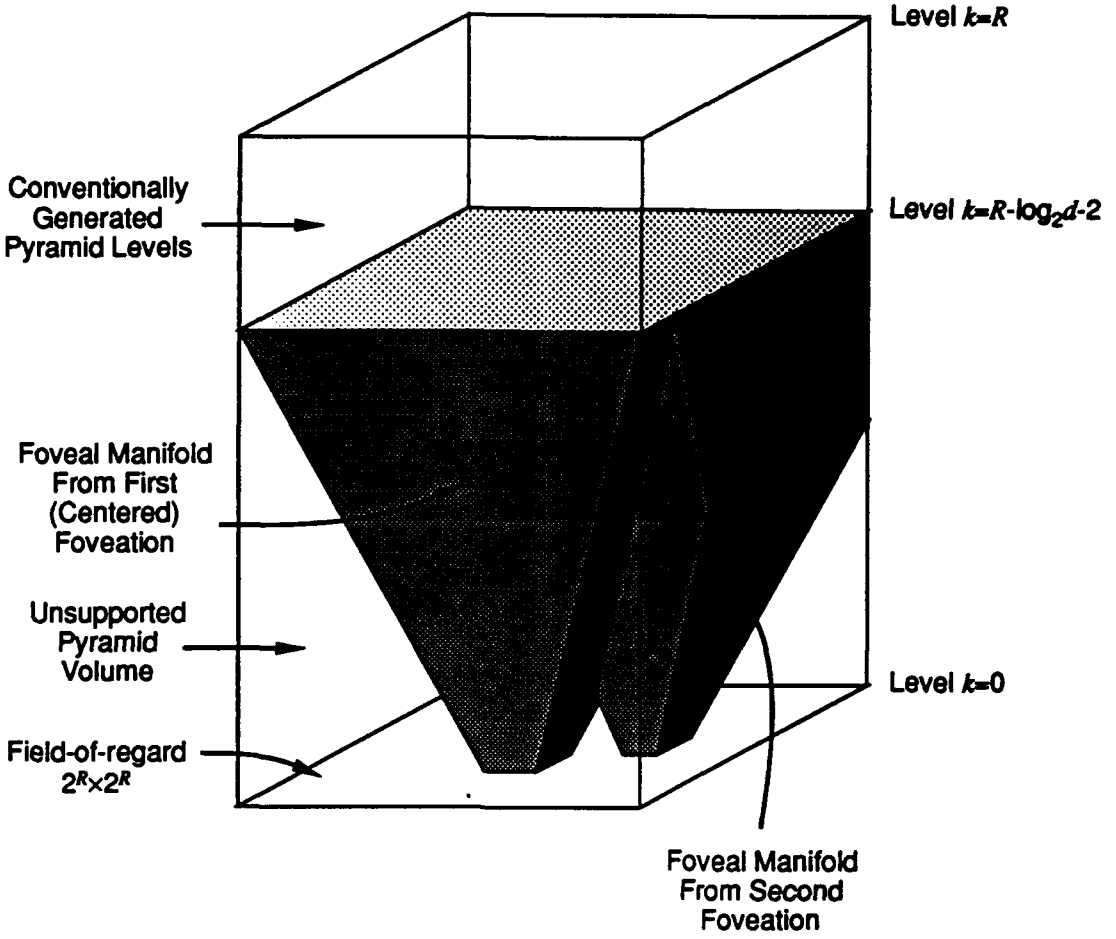


Figure 6.2.4-1. Foveal polygon after two foveations.

Spatial miscorrelation can be controlled at the expense of constrained foveation locations. Specifically, if the optical axes associated with a set of foveations fall on the center of cells at at some level k, then the frames will produce cells that map directly to levels 0 through k. For a bottom-up generation process with a=m=2, the number of cells at level k is a factor 4^{-k} less than at level k=0, and so the possible foveation locations preserving correlation at level k is likewise reduced from the total number of possible locations. The foveal axis locations must satisfy

$$\left(x_{c,k}, y_{c,k}\right) = \left(i2^k, j2^k\right) \tag{6-17}$$

where k is the highest polygon level where correlation must be preserved (determined by the vision task and the task state), i and j are integers in [0...$\log_2 N$], and the field-of-regard is N×N. The effect of this constraint is minimized by requiring correlation at the lowest possible levels in the polygon. The effect of cell interpolation to correct for

miscorrelation (interpolation error) is less significant at higher levels since the direct mapped values are themselves averages across relatively large regions. Levels at and above the top of the foveal manifold produced by a foveation to the center of the field-of-regard (typically the first registration) need not be overwritten, since they are completely specified.

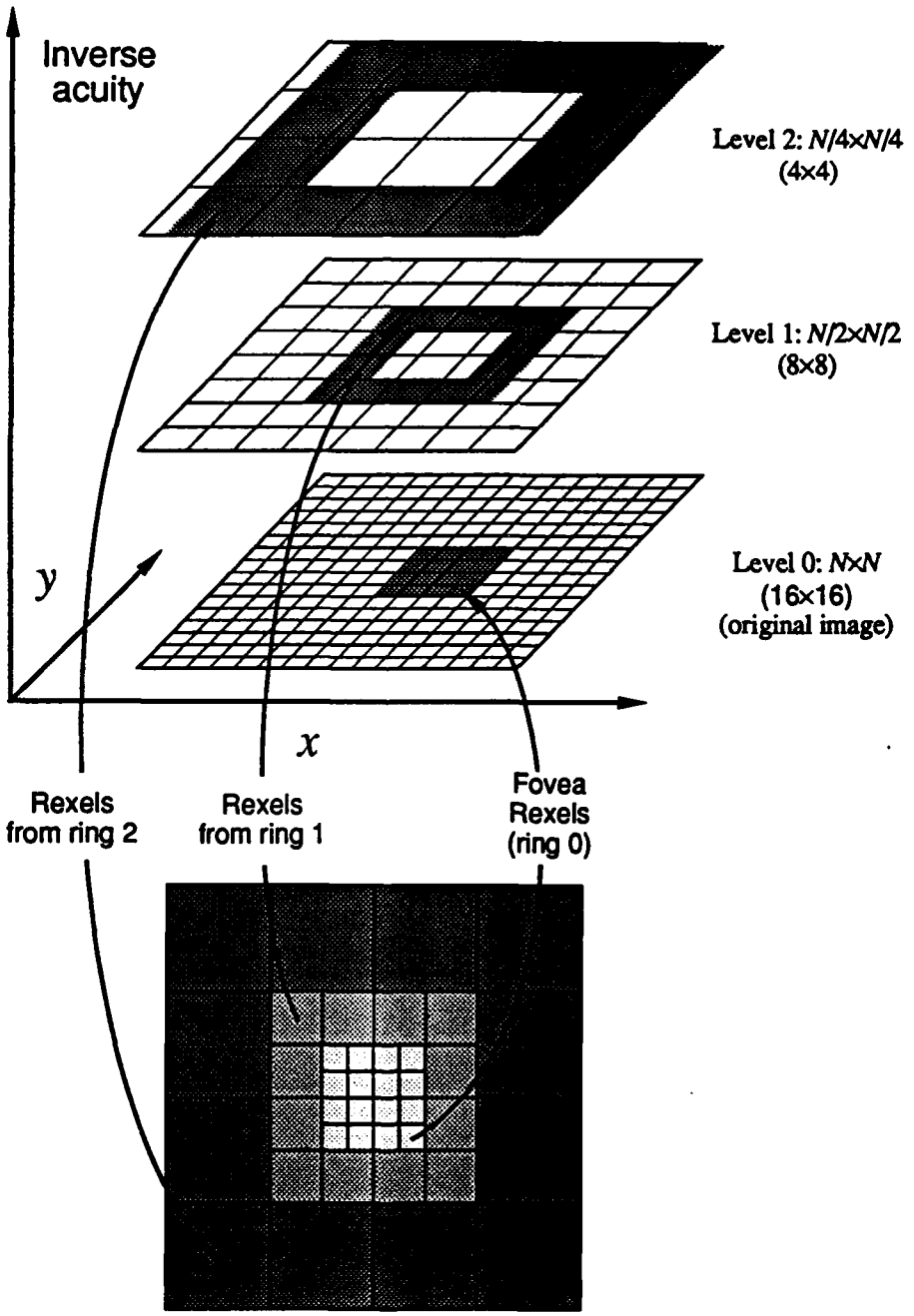


Figure 6.2.4-2. Spatial miscorrelation between rexels and pyramid cells.

When spatial miscorrelation exists between rexels and the cells of the polygon, interpolated values based on the rexel data can be stored in the polygon. No miscorrelation ever exists in the base because all locations (within the maximum acuity of the system) are supported. However, there can be a 50% offset between rexels and cells at the second level (k=1), as shown in Figure 6.2.4-3. In general, the offset between rexels and cells at some level k is any integer multiple of $2^{-k}\times100\%$ between 0% and 50%. The value for ${}^1G_{3,4}$ is computed by averaging the central four rexels of the fovea. The value for ${}^1G_{2,5}$, however, is the average of the top left rexel of the fovea and top left three 2×2 rexels of the second ring.

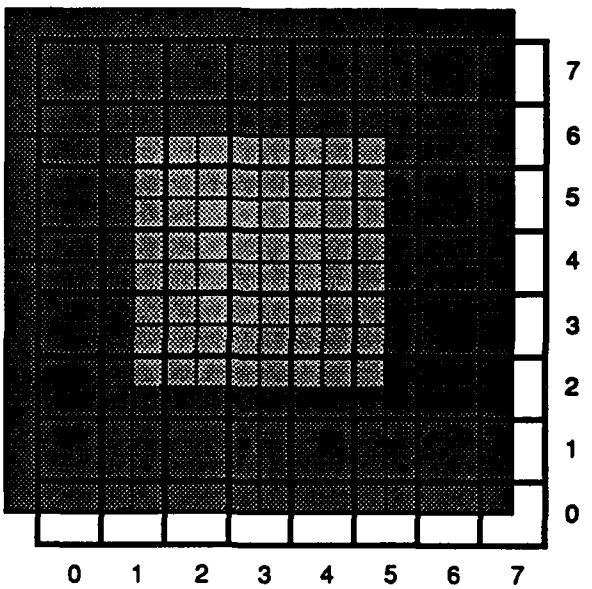


Figure 6.2.4-3. Spatial miscorrelation between rexels and polygon cells at level k=1.

In a certain sense it is misleading to store this interpolated value in level k=1 because it is an average over an area larger than 2×2. Following the discard method of integrated perception generation, when there exists for a given location in the field-of-regard two rexels of different acuity, the higher acuity rexel is retained. When information is represented in three dimensions (versus two), the possibility for a "collision" is greatly reduced. Specifically, the values of both rexels (or interpolated values) can be retained in the polygon at similar spatial location but different acuity levels, and collision occurs only when rexels of the same size overlap.

6.3 Image Processing and Foveation Strategies using the Foveal Polygon

An important feature of the foveal polygon is that it supports conventional pyramid algorithms and (at the individual levels) the uniform resolution image processing of foveal data. Even though an entire pyramid data structure cannot be built from a frame of foveal data, the subset it does represent can be processed conventionally. Indeed, the premise of foveal vision is that scene information relevant to the vision task is localized, and only these regions need to be resolved.

A second feature of the foveal polygon is that it supports dynamic closed loop foveation strategies in a very straightforward fashion. Specifically, as hierarchical top-down analysis algorithms process a foveal polygon, they can also indicate where the system should look, and even how closely (e.g., a direct foveation to location *A*, or foveation "in the vicinity" of location *B*). This information is used by gaze control algorithms to select the sensor gaze angle. Constraints can be imposed on gaze control, such as minimizing the number of foveations or overall sensor movement.

A top-down analysis can be considered as an interrogation process which, starting with a low acuity representation of the scene, formulates hypotheses based on detected cues (features) and confirms or denies the hypotheses by interrogating (analyzing) the associated cues with higher acuity. Any number of these interrogations can require data outside the polygon, or equivalently, information on some particular region in the scene at higher resolution than that provided by the foveal frame. Such will often be the case when analyzing a small or detailed (high bandwidth) feature in the periphery of the field-of-view. These requests define *candidate locations* for foveations.

When the required resolution at a candidate location is less than the maximum foveal sensor resolution (i.e., the resolution at the fovea), then it may not be necessary to point the optical axis directly at that location. This is important in scenes with multiple features (as is typically the case) or when performing parallel tasks, because it allows gaze angles to be computed which can satisfy the resolution requirements of several candidate locations simultaneously (e.g., foveating to the centroid of a feature cluster). Specifically, if multiple regions have to be further resolved, each by some proposed amount (measured in acuity or polygon levels), then a minimum length sequence of optical axis locations can be selected which satisfies the resolution requirements. As in nature, required information

on features can be obtained in many cases without foveating directly at each one [Berge83]. Thus, a more efficient allocation of acuity is achieved.

When a top-down pyramid algorithm is analyzing a scene feature, an estimate can often be formulated on the acuity necessary to complete the analysis. Scale invariance is obtained when necessary acuity is defined in terms of levels below that in which the object is first detected. For example, the segmentation of an object to its components may require five levels below that in which the object first appears in an unresolved state. The number of levels can be obtained from a stored model of the object. This is not to say that the estimate is deterministic; upon resolving an object, it may turn out to be a totally different object than that hypothesized, requiring fewer or more levels to segment. Nevertheless, the estimate of required resolution computed by top-down algorithms processing polygon data, in conjunction with the sensor acuity profile, can be used as a measure of how close the foveal sensor must look at the object.

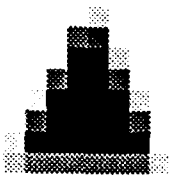
As an example, consider a simple classification task. The scene consists of an object known to be either a triangle or a star, and the objective of the system is to classify the object into one of these two categories. A hierarchical model of the geometric shapes shall be used in which the shape is represented at different resolutions, or levels (Figure 7.3-1). Obviously, templates for geometric shapes can be represented with infinite acuity; a maximum acuity is defined which resolves the shape sufficiently to serve all the interests of the vision system. This level is not necessarily the pixel level; if n connected cells define a star with acceptable resolution, then it should not matter if the cells are pixels forming a small star or higher level ($k>0$) cells forming a large star.

At level $k_m=3$ and higher, the representations of the star and triangle may be difficult to distinguish (e.g., with matched filters). This is especially true if the image is thresholded (a common intermediate step in object analysis). If the foveal sensor registers the object at an acuity equal to or less than that of level $k_m=3$, then the system may have to obtain higher acuity data on the object by registering another frame of sensor data with the optical axis closer to the object.

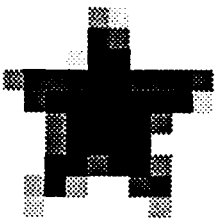
Let a top-down analysis of the top few levels in the foveal polygon discover an unresolved object at level $k=s$, and label it as an object for further interrogation. This object representation at polygon level $k=s$ can be similar (i.e., area in cells) to the object representation at model level $k_m=5$. Furthermore, let the resolution at model level $k_m=2$ be the minimum resolution supporting the confidence necessary for final classification. Thus,

a. Base of triangle model.

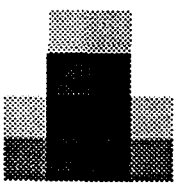
b. Base of star model.



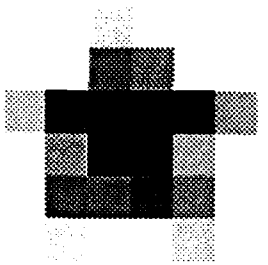
c. Level k_m=2 of triangle model.



d. Level k_m=2 of star model.



e. Level k_m=3 of triangle model.



f. Level k_m=3 of star model.

g. Thesholded (50%) level k_m=3 of triangle model.

h. Thesholded (50%) level k_m=3 of star model.

Figure 6.3-1. Hierarchical representations of geometric shapes. Each model consists of a six level (k_m=0,...,5) Gaussian pyramid generated from a pixel representation of the shape roughly centered at the base. At level k_m=3, the shape is nearly unresolved.

to accomplish the vision task, the object must be represented within the foveal polygon at the polygon level k=s-5 or lower. This is satisfied as long as the object is registered by any of the sensor's first s-5 rings. Of course, if s<5, then the necessary acuity for classification with acceptable confidence exceeds that of the machine vision system (if the system were mobile, one recourse would be to move closer).

Minimum object resolution requirements can be estimated from more heuristic means when an object model is not well defined or available. One such straightforward approach is to define some minimum number of cells N_c with which an object should be resolved before analysis results can be performed with confidence. Consider some object resolved at polygon level k to $N_{o,k}$ cells, i.e., there are $N_{o,k}$ connected cells in a thresholded representation of the level. Each cell in the interior of the object at level k will have four corresponding cells in the representation of the object at the resolution of level k-1 (this is not to say that level k-1 of the polygon will register the entire object). The perimeter cells in level k will have from one to four corresponding cells at this higher resolution. For even small regions, the number of interior cells is significantly greater than the number of peripheral cells, so a quadruple increase in object area (measured in cells) for a doubling of representation resolution can be used as a rough guideline. We shall call this the *power-of-four rule.*

Table 6.3-1 presents the areas of triangle and star objects at the different levels of their model, and shows how the areas of the thresholded objects (such as in Figure 6.3-1g, h) conform roughly with this power-of-four rule. Area estimates are extrapolated from the level in which object presence would first be hypothesized, e.g., when a small group of a few connected thresholded cells appear, to illustrate worst case performance.

Hierarchical Model Level	Area of Triangle (Thresholded)	$4^{5-k} \times N_{o,k}$ for Triangle	Area of Star (Thresholded)	$4^{5-k} \times N_{o,k}$ for Star
5	6	6	7	7
4	18	24	31	28
3	92	96	118	112
2	341	384	480	448
1	1338	1536	1890	1792

Table 6.3-1. Extrapolated (power-of-four rule) and actual areas of objects.

The heuristic power-of-four rule predicts that an object resolved at level k with $N_{o,k}$ cells, $N_{o,k} \leq N_c$, should be resolved by at least N_c cells at level $k-\lceil \Delta_l \rceil$, where Δ_l is given by

$$\Delta_l = \log_4\left(\frac{N_c}{N_{o,l}}\right) \tag{6-18}$$

If the foveal polygon does not support analysis of the object at levels below k, then $k-\lceil \Delta_l \rceil$ is a good measure of how close to the object (in major rings) the sensor optical axis must be situated in order to obtain the data necessary to continue with the top-down analysis of the object. A map of the field-of-regard can be generated with the object location labeled with the "request for servicing" by ring $k-\lceil \Delta_l \rceil$ or better. This list of locations and corresponding desired resolutions is called the *service map*. A straightforward foveation control strategy is one which inspects the service map after top-down analysis can no longer continue with the data in the polygon, and services the request for additional information. When there is only a single request (e.g., only one object in the scene, or multiple objects with all but one sufficiently well supported by obtained information), the strategy is trivial: foveate directly to the location of the service request.

The top-down analysis of the polygon data generated from a registration of a complex scene can detect several cues and attempt to process them. When these cues are to be analyzed with high resolution, and are widely dispersed with respect to the area of the fovea, it is unlikely that the polygon will contain all the information required by the top-down analyses to completely process all the cues. In this case, the service map will contain multiple requests, situated in different locations within the field-of-regard and requesting different resolutions (Figure 6.3-2). The saccadic foveation strategy must service these requests with the minimum number of foveations so as to minimize the amount of overall data generated and processed by the vision system. The two types of foveation strategies discussed in Chapter 5, survey mode and interrogation mode, apply to the foveal image processing of resolved objects. Survey mode attempts to resolve request clusters on the service map by foveating to cluster centroids. Interrogation mode attempts to align the optical axis directly on some particular feature being interrogated.

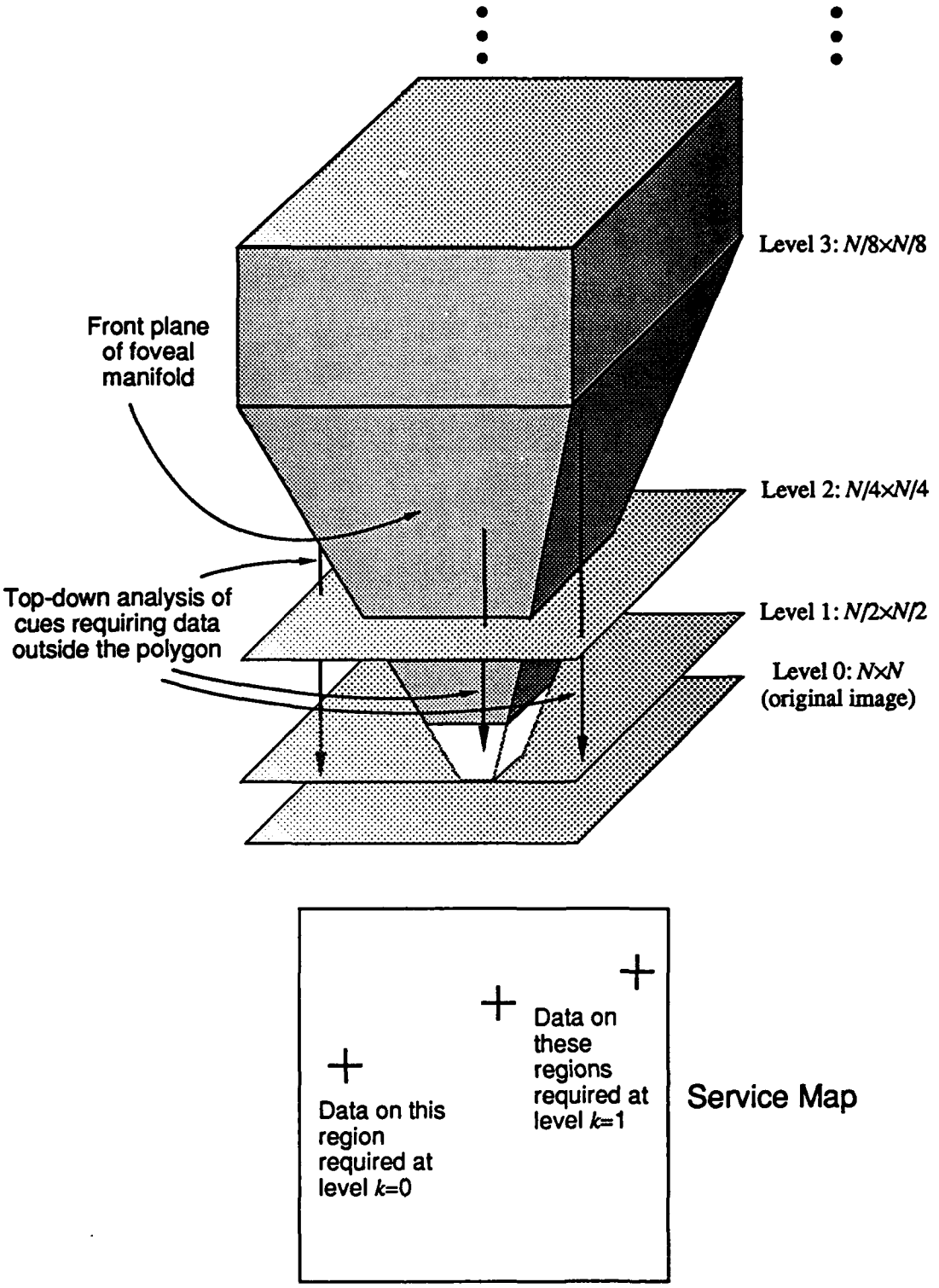


Figure 6.3-2. Foveation control strategy service map. Each entry in the map represents a location where top-down analysis requires additional data and the necessary acuity.

6.4 Foveal Image Processing Exercises

This section presents three exercises of a foveal machine vision system. Several foveal data processing techniques and foveation strategies are discussed. The first exercise forms a Gaussian polygon from the integrated perception generated by the discard method. Conventional area analysis and edge enhancement techniques are used in the top-down algorithm to distinguish objects in the scene. A survey mode gaze control strategy is employed. The second exercise uses both a Gaussian and Laplacian polygon. Sensor frames are processed separately, and a low level integrated perception is not formed. Area analysis techniques are used for preliminary object discrimination, and template matching is used to interrogate detailed features. An interrogation mode strategy is employed. The third exercise is a series of simulations using variations of the scene from the second exercise. For all exercises, an exponential pattern uniformly subdivided by a factor of 4 as illustrated in Figure 3.5-1 is used, with a field-of-view of 512×512 pixels for the first two exercises, and 256×256 pixels for the third. The field-of-regard is 512×512 pixels. Of significance is that the image processing algorithms used in the exercises are scale and rotation invariant.

The objective of this section is to illustrate in operation the concepts discussed thus far. Emphasis is placed on gaze control and the sequence of events, and not on details of the image processing algorithms. These are treated in greater detail in other sources dedicated to pyramidal and other forms of hierarchical image processing [Bessl86], [Clark84], [Crowl84], [Dyer87], [Grosk84], [Miller88$_1$], [Miller88$_2$], [Shein84], [Stout86], [Stout87], [Tanim84], [Zucker84].

6.4.1 Counting Pennies Distributed Among Other Objects

In this exercise, the foveal machine vision system is tasked with resolving a number of different objects situated such that multiple foveations are necessary. The exercise combines integrated perception generation from foveal data and a service map directed gaze

control strategy with conventional shift invariant image processing techniques, specifically area analysis (area and aspect ratio measurements).

The scene, shown in Figure 6.4.1-1, consists of a number of pennies and paper clips. The objective of the machine vision system is to count the pennies. Object cells are initially distinguished (labeled) from background cells by grey value thresholding at the lowest level in the polygon in which the entire field-of-regard is supported, i.e., the waist. There are 16×16 cells at this level, so a cue (group of connected cells labeled as object) can be detected while processing a relatively small amount of data.

To reduce the possibility of perceiving multiple objects as one cue, the level is convolved with a simple edge enhancement kernel (Figure 6.4.1-2) and histogram equalized prior to thresholding. The filtering operation attenuates smooth regions between strong cues, thus enhancing their localization. In effect, the highpass filtering compensates, to a limited extent, for the blurring of features in the low acuity representation. After histogram equalization, the waist is thresholded at 50% to produce a binary image which labels the cells; cells with grey values below the threshold are labeled as object cells. The filtering and equalization are instrumental in keeping the values of cells which bridge different cues to above the 50% threshold.

A top-down algorithm will discriminate and label cues by their area and aspect ratio as follows:

1. Resolve the cue to a minimum of 18 cells (thresholded).

2. Compute the major and minor axes of the ellipse that best fits over the thresholded object. Label the object as a penny if the aspect ratio of the object is within 1:1±0.2 (this range accounts for the variability in aspect ratio of circles represented with 18 cells)

The principle role of the foveation control strategy is to provide the algorithm with the measurement data required to accomplish step 1. The power-of-four rule will be used to predict the highest polygon level which resolves each cue to at least 18 cells. A service map will be formed, and the foveation strategy will attempt to satisfy the minimum cue resolution requirements with as few registrations as possible by servicing multiple cues with one foveation.

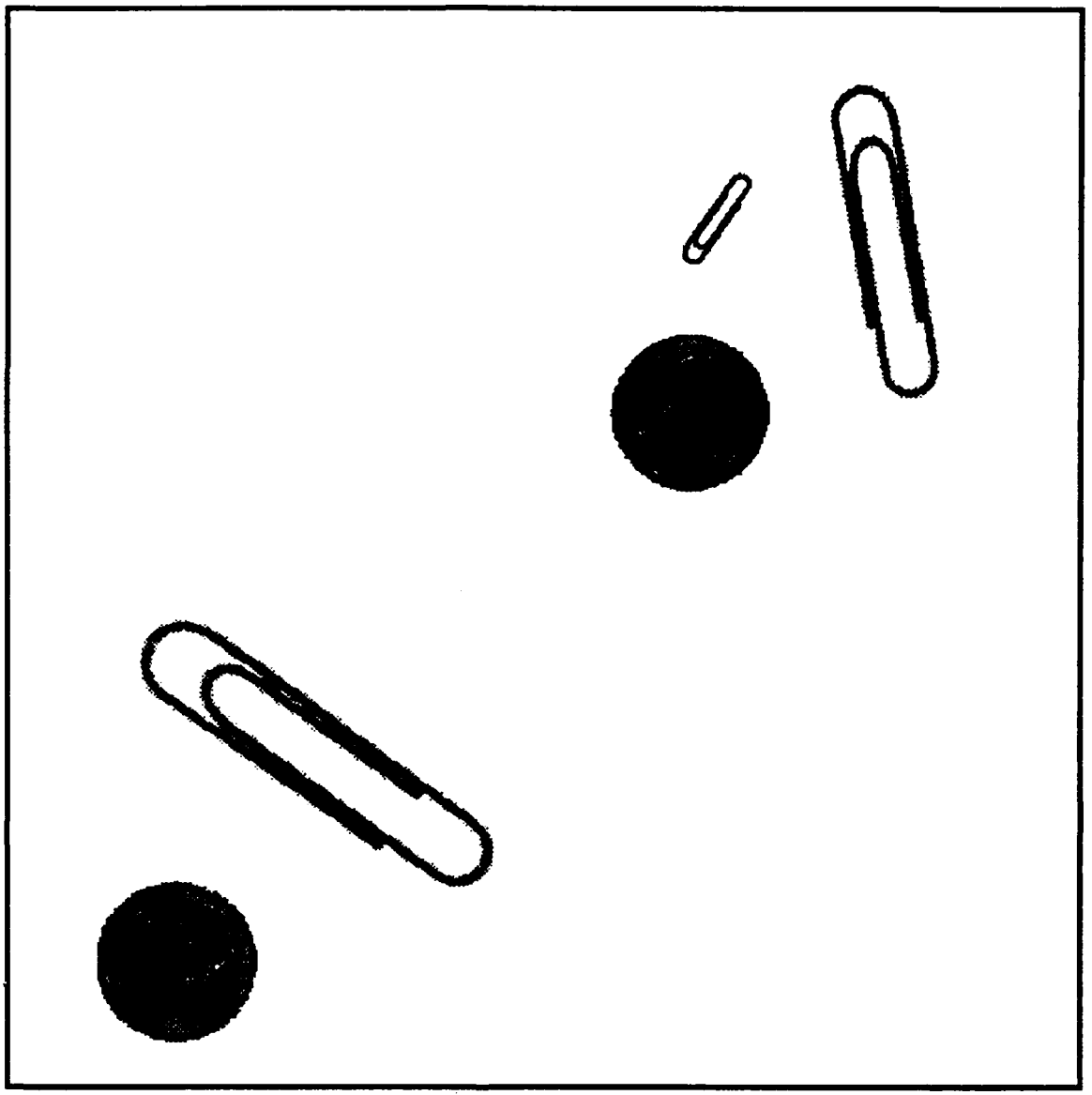


Figure 6.4.1-1. Pennies in clutter scene. The scene consists of 512×512 pixels.

-1	-1	-1
-1	9	-1
-1	-1	-1

Figure 6.4.1-2. Highpass filter kernel. This filtering operator localizes cues.

The initial registration is made with the optical axis centered over the scene (Figure 6.4.1-3), generating 1216 rexel values. A Gaussian polygon with ten levels (k=0...9) is generated. The top of the foveal manifold is at level k=5 of the resulting polygon, and consists of 16×16 cells (Figure 6.4.1-4).

After the highpass filtering, histogram equalization, and thresholding of level k=5, five cues are identified (Figure 6.4.1-5) which are now analyzed by the top-down algorithm. The value of this conditioning step can be appreciated by observing how the cues are defined when polygon level k=5 is thresholded directly, in which case most cues are missed, or when the level is only histogram equalized prior to thresholding, increasing the susceptibility of the initial cue labeling process to cue connecting cells (Figure 6.4.1-6). Depending on the specific labeling algorithm, Figure 6.4.1-6b would register the five scene objects as 2, 3, or 4 cues.

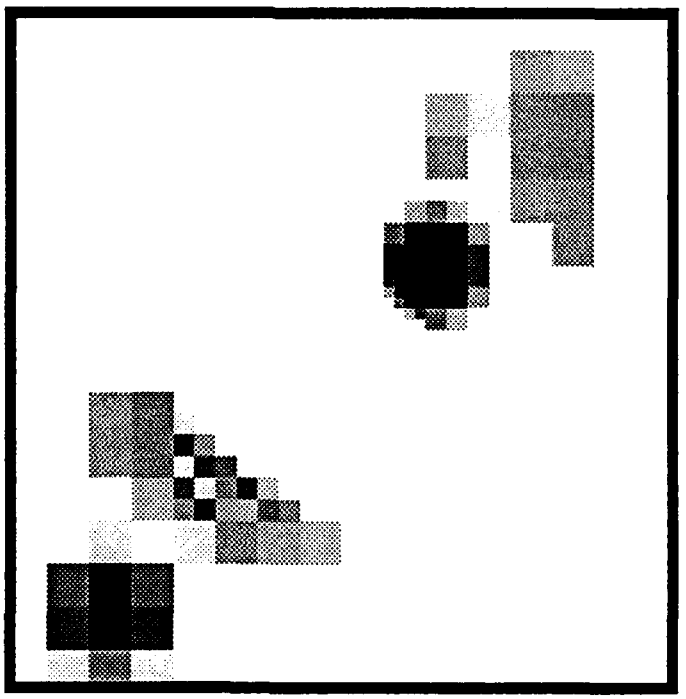


Figure 6.4.1-3. First foveal sensor frame.

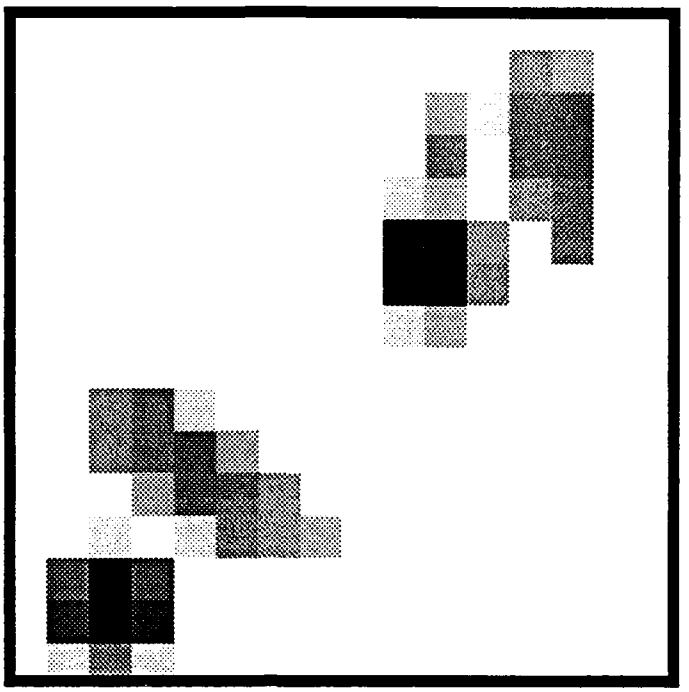


Figure 6.4.1-4. Top level of foveal manifold (k=5) representing the first sensor frame.

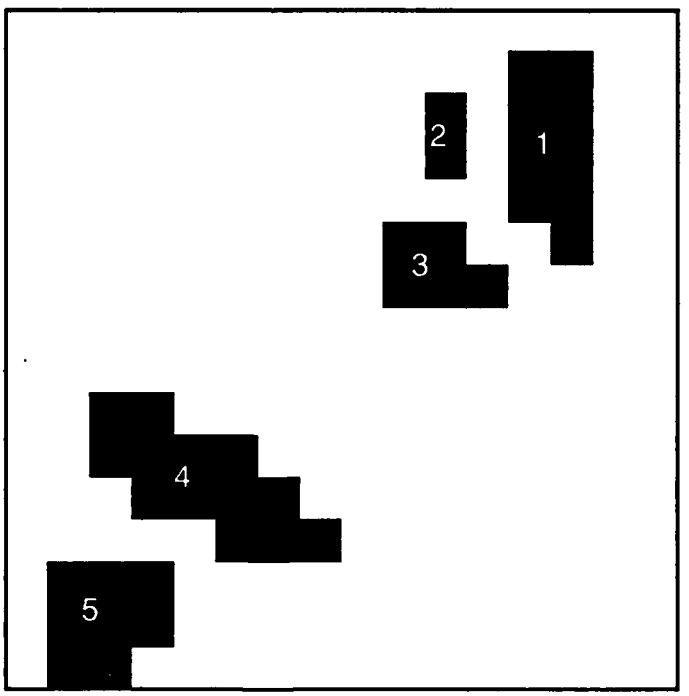


Figure 6.4.1-5. Cue labeling of conditioned and thresholded polygon level k=5.

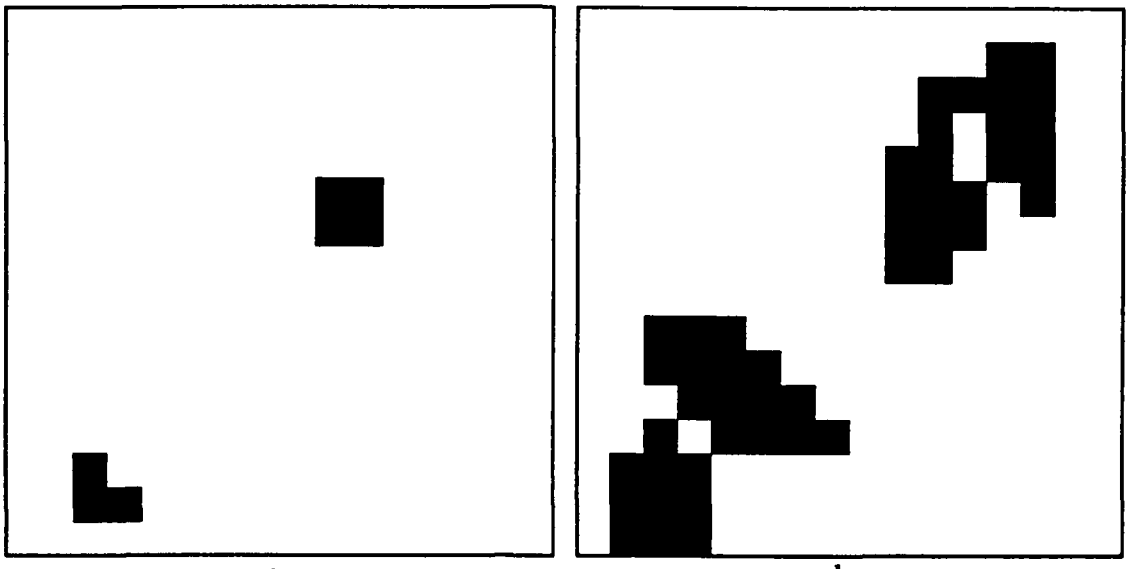


Figure 6.4.1-6. Cue labeling without image conditioning. (a) Direct thresholding of level k=5. (b) Thresholding after histogram equalization highpass without filtering.

The initial statistics on the cues are presented in Table 6.4.1-1. From this information, a service map for the gaze control strategy is generated in a straightforward fashion. Equation (6-18) provides an estimate for the levels in which the five cues will be resolved with at least 18 cells. These estimates are given in Table 6.4.1-2 and shown graphically in the format of a service map in Figure 6.4.1-7. Here, the service map consists only of the centroid location and desired acuity for each cue. Also shown in Figure 6.4.1-7 is the acuity supported by the foveal sensor.

Cue	Area $N_{o,5}$ (in thresholded cells)	x coordinate of cue centroid	y coordinate of cue centroid
1	9	12.5	12.0
2	2	10.0	12.5
3	5	10.0	9.5
4	13	4.5	4.5
5	8	2.0	1.0

Table 6.4.1-1. Initial cue statistics. Locations are given in terms of level k=5 coordinates.

Cue	Δ_l $\log_4(18)-\log_4(N_{o,5})$	$INT\uparrow[\Delta_l]$	expected level with sufficient resolution
1	0.50	1	4
2	1.58	2	3
3	0.92	1	4
4	0.23	1	4
5	0.58	1	4

Table 6.4.1-2. Estimates for cue resolving polygon levels.

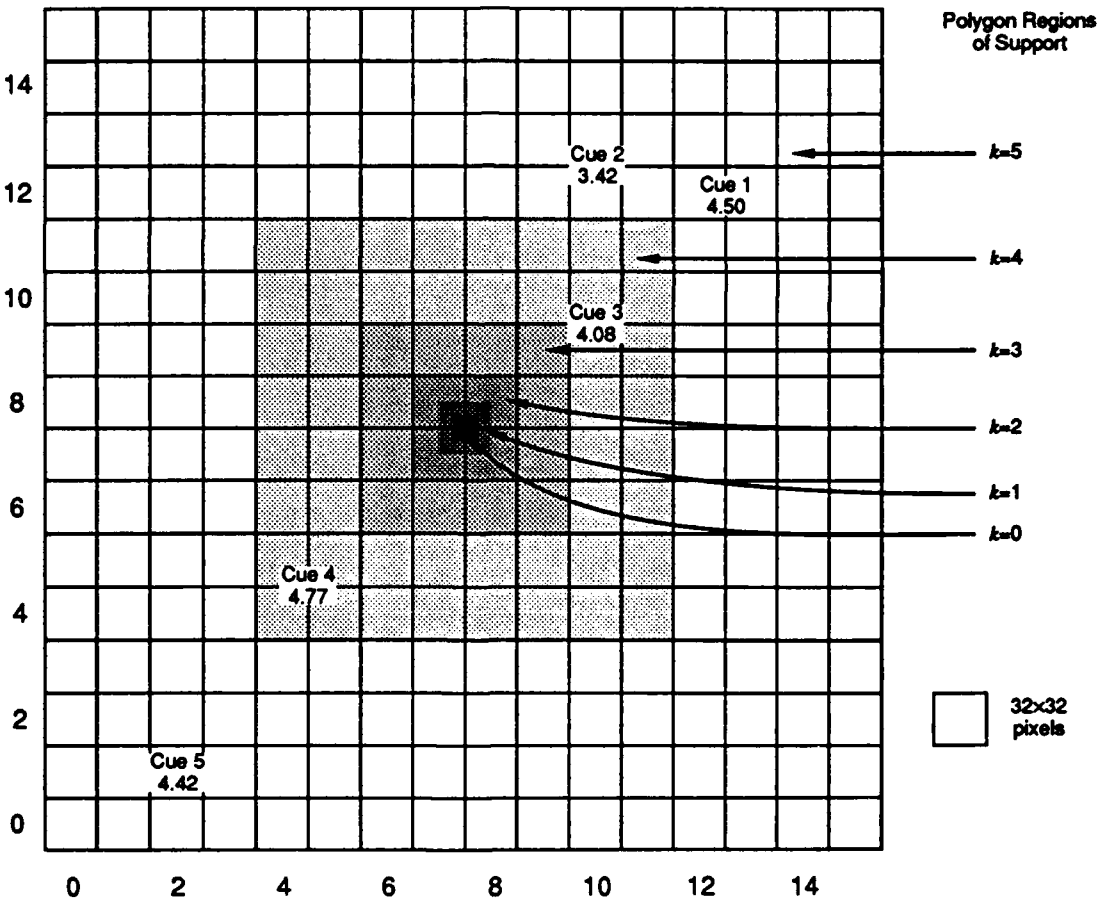


Figure 6.4.1-7. Service map after processing first sensor frame. Locations of cues and estimated necessary acuity (in terms of polygon level) are shown over sensor acuity. Cell coordinates correspond to those of polygon level k=5.

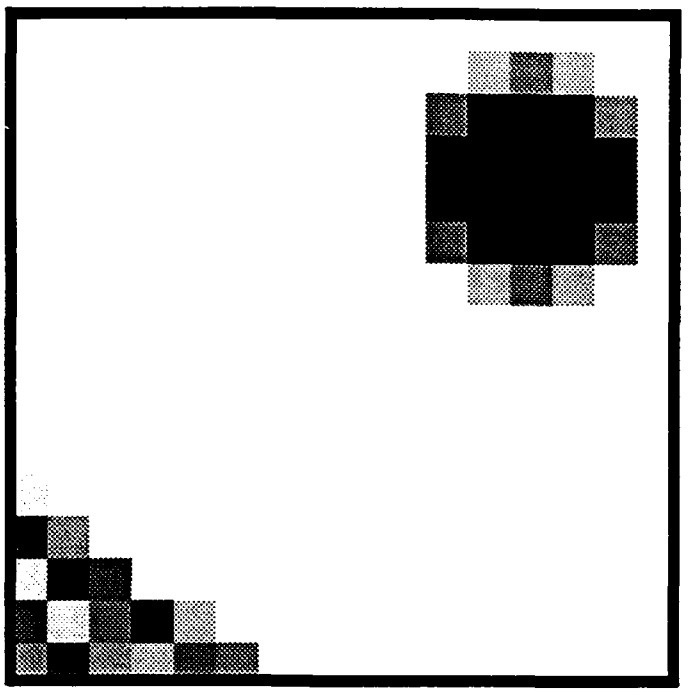
All the cues require greater resolution than that provided by polygon level k=5. The centroids of cues 1, 2, and 5 fall on the region of the scene supported by the sensor ring corresponding to level k=5. Therefore, there is no additional (higher acuity) data on these three cues (they are supported by the wide field-of-view low acuity perception but not by any lower field-of-view higher acuity perception).

Cues 3 and 4 are predicted to require the degree of acuity provided by level k=4. The centroid of cue 4 falls in this region of the first sensor frame, but because of the object orientation, the cue area is predominantly defined by level k=5 data. Additional data at level k=4 (or better) will be required from the following foveations to resolve the cue. It is seen that for semiresolved objects, resolving the vicinity of the cue centroid does not imply resolving the entire cue. Figure 6.4.1-8 shows level k=4 of the polygon for the first sensor frame, and how only the central portion of cue 4 is resolved with this degree of acuity. One test to check the completeness of a cue representation at some level is to confirm that no thresholded cells about the centroid touch the perimeter of the level. The drawback of this approach is that it reduces the effective coverage of the polygon level by two rings of cells, or from $4d \times 4d$ to $(4d\text{-}2)\times(4d\text{-}2)$. Another approach is to label the service map with the entire cue area, and not just the centroid.

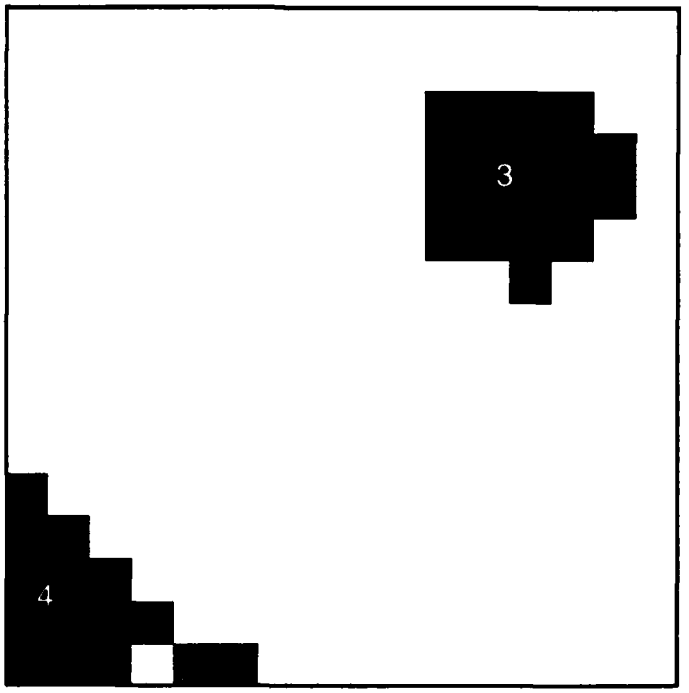
The region of cue 3 (its area in k=5 level cells) is supported entirely at level k=4 as shown in Figure 6.4.1-8. At this level, cue 3 is resolved to 19 cells, meeting the requirement for further analysis. The best fitting ellipse over cue 3 has a major axis of length 2.57 cells making an angle of 173° with the horizontal, and a minor axis of length 2.36. The aspect ratio of the ellipse is 1:1.09, so the cue is classified as a penny.

Two foveations are identified in the gaze control service map as required to resolve the remaining cues: one in the vicinity of cues 1 and 2, and another in the vicinity of cues 4 and 5. The first foveation will be to the centroid of the cluster cue 1 and cue 2. The selection of the optical axis location (x_c,y_c) when resolving a cluster should be weighted by the resolution requirements of the individual cues so as to bias the axis to the cues that need to be more resolved. A simple linear bias will be employed here:

$$x_c = \frac{\sum_{i=1}^{n} x_{c,i}\Delta_i}{\sum_{i=1}^{n} \Delta_i} \qquad (6\text{-}19)$$



a.



b.

Figure 6.4.1-8. Level k=4 of polygon generated from first sensor frame. (a) Actual level, and (b) after conditioning and labeling

$$y_c = \frac{\sum_{i=1}^{n} y_{c,i}\Delta_i}{\sum_{i=1}^{n} \Delta_i} \tag{6-20}$$

where n is the number of cues in the cluster, $(x_{c,i},y_{c,i})$ is the center of the i'th cue in the cluster, i=1...n, and Δ_i is the required improvement in resolution for the i'th cue. Table 6.4.1-1 gives the location of cue centroids in terms of level k=5 coordinates. The relationship between level k_1 coordinates $(x_{k_1}y_{k_1})$ and level k_2 coordinates $(x_{k_2}y_{k_2})$, $k_1 \le k_2$, is

$$x_{k_2} = 2^{k_2-k_1} x_{k_1} \tag{6-21}$$

$$y_{k_2} = 2^{k_2-k_1} y_{k_1} \tag{6-22}$$

From (6-19) through (6-22), the axis for the foveation to the first cluster is (339.2,396.2) in k=0 coordinates. In this exercise, an integrated perception will be formed using the discard method such that no cell interpolation will be required at or below level k=4. Note that since at k=5 the entire field-of-regard is addressed by the first foveation, no cells at this level will be replaced. To avoid the cell interpolation, the axis location will be centered at one of the cells at k=4, forcing the axis to be at some location defined by (6-17). The axis location is thus "quantized" to the value (21,25) in level k=4 coordinates, or (336,400) in k=0 coordinates.

The sensor frame obtained by the second registration in the simulation is given in Figure 6.4.1-9, and the integrated perception is represented in Figure 6.4.1-10 in the format of Chapter 4. The second frame consists of 1151 rexels falling in the field-of-regard, and 62 rexels of the first frame are replaced by 952 rexels of the second frame in the integrated perception which now consists of 2106 rexels. Levels k=0 through 4 no longer consist of simple square regions, because of the contribution of data by the first and second sensor frames at different location. Figure 6.4.1-11 illustrates the coverage of the first frame (centered border) and the second frame (upper right hand border) at level k=4 of a Gaussian pyramid formed by a uniformly sampled pixel registration of the scene. The Gaussian polygon at level k=4 after the fusing of the two foveal frames consists of the union of the two bordered regions. Note that the region of support at k=4 of the second frame is not square. Some of the top rexels fall out of the field-of-regard, cropping the normally square region of sensor frame support (in this case 16×16 cells).

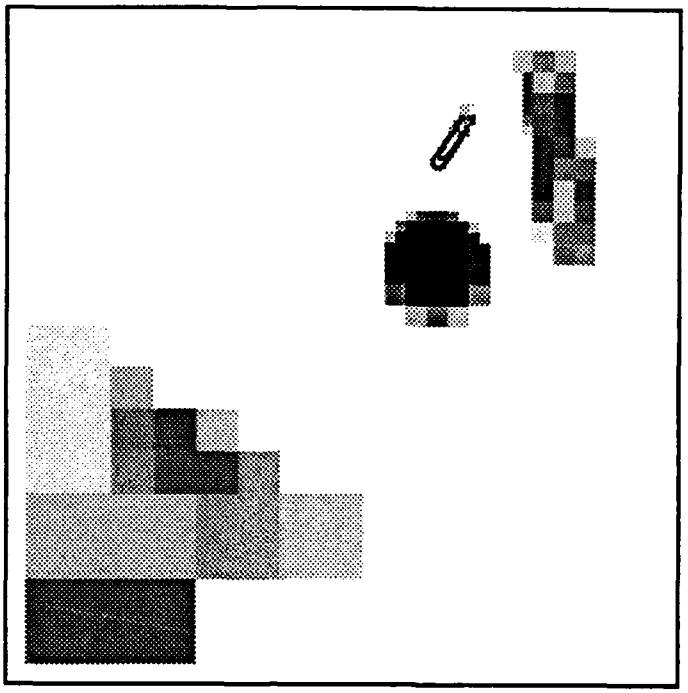


Figure 6.4.1-9. Second foveal sensor frame.

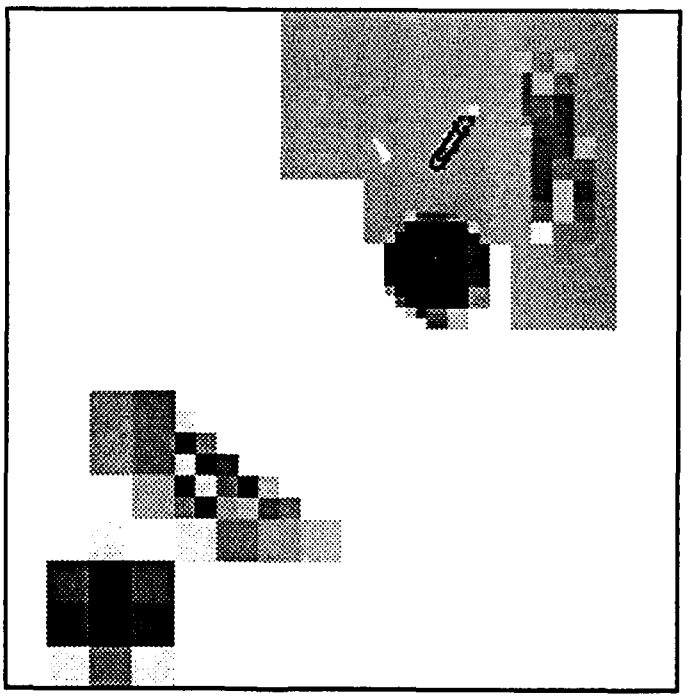


Figure 6.4.1-10. Integrated perception of first two registrations using discard method. The light gray region on the top represents the contribution to the first integrated perception by the second sensor frame.

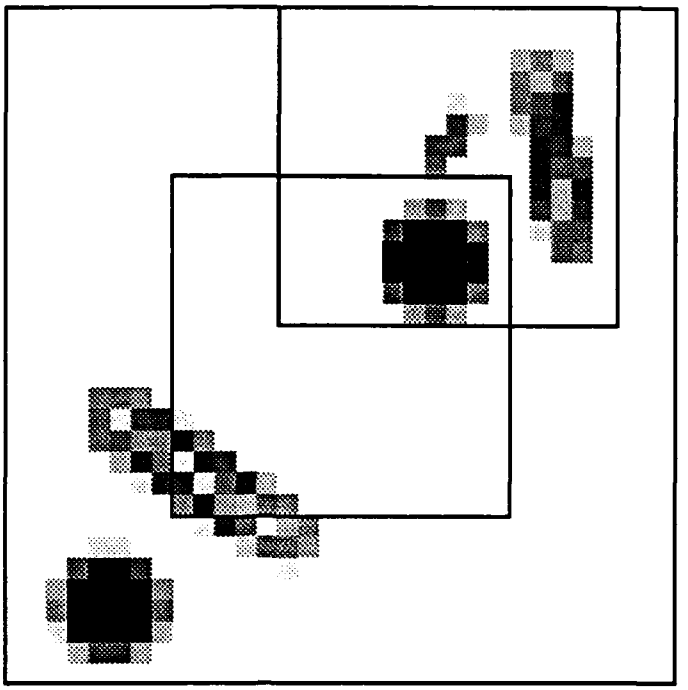


Figure 6.4.1-11. Coverage within a complete pyramid level k=4 by data from the first and second sensor frames.

Level k=4 of the new polygon is shown after conditioning and cue labeling in Figure 6.4.1-12 with respect to the coordinates of pyramid level k=4. Cue 1 is resolved to 23 cells, and the best fitting ellipse has a major axis of length 5.7 cells making an angle of 145° with the horizontal, and a minor axis of length 1.45. The aspect ratio of the ellipse is 1:3.94, so the cue is classified as not a penny.

Cue 2 is resolved in level k=4 to 5 cells, less than the 18 necessary to conduct the aspect ratio test. This is to be expected, since the prediction made from level k=5 was that k=3 would be the highest level adequately resolving the cue. Figure 6.4.1-13 illustrates the coverage of the first frame (centered border) and the second frame (upper right hand border) at level k=3 of the polygon with respect to the coordinates of pyramid level k=3. At this higher acuity level, the coverage of the frames is smaller and disjoint (Figure 6.4.1-12 represents a slice of the legs of the two frame polygon illustrated in Figure 6.2.4-3). Level k=3 of the new polygon is shown after conditioning and cue labeling in Figure 6.4.1-14 with respect to the coverage of the second frame at that level. Cue 2 is resolved at k=3 to 19 cells, and the best fitting ellipse has a major axis of length 3.98 cells making an angle of 58° with the horizontal, and a minor axis of length 1.52. The aspect ratio of the ellipse is 1:2.6, so the cue is classified as not a penny.

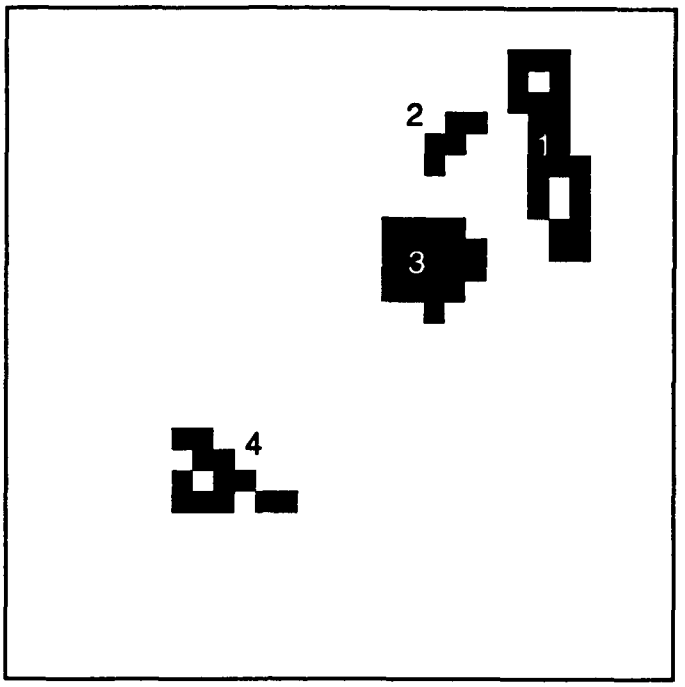


Figure 6.4.1-12. Conditioned and cue labeled polygon level k=4 of first two frames. The image is in the coordinates of pyramid level k=4.

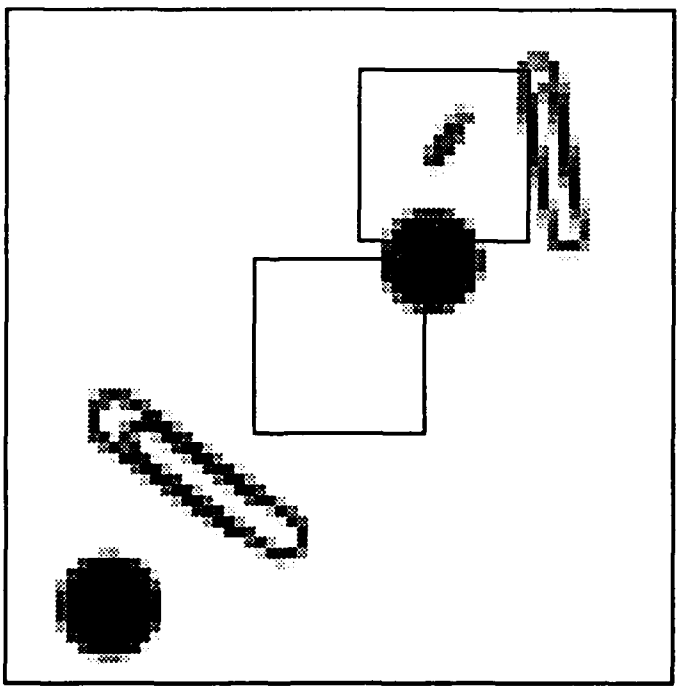


Figure 6.4.1-13. Coverage within a complete pyramid level k=3 by data from the first and second sensor frames.

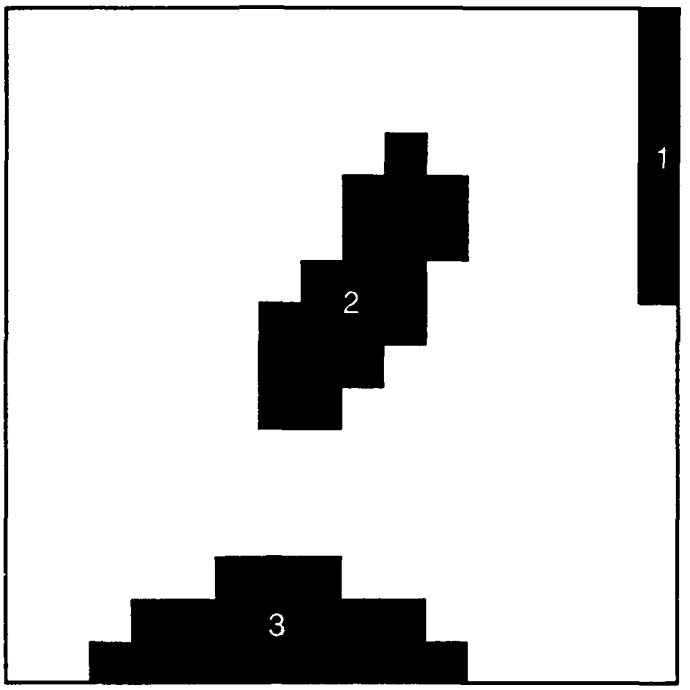


Figure 6.4.1-14. Conditioned and cue labeled polygon level k=3 of second frame. The image is in the coordinates of the region of coverage of the second frame at k=3.

All the cues in the first cluster have been processed, and the system now proceeds to resolve the second cluster (cues 4 and 5). From (6-19) through (6-22), the axis for the foveation to the first cluster is (86.7,63.8) in k=0 coordinates. Quantizing the location to avoid miscorrelation with the existing data at level k=4 gives (80,64). The third sensor frame is given in Figure 6.4.1-15, and the integrated perception is represented in Figure 6.4.1-16. The third frame consists of 1034 rexels falling in the field-of-regard, and 34 rexels of the first frame are replaced by 904 rexels of the third frame in the integrated perception which now consists of 2976 rexels. Figure 6.4.1-17 illustrates the coverage of the three frames with respect to level k=4 of a conventional pyramid. As with the second frame, some rexels of the fifth ring (level k=4 acuity) fall outside the field-of-regard.

The conditioned and labeled level k=4 of the new polygon is shown in Figure 6.4.1-18 with respect to the coverage of the third frame at that level. The majority of the level k=5 area of cue 4 is resolved to 27 cells, and the best fitting ellipse has a major axis of length 6.2 cells making an angle of 146° with the horizontal, and a minor axis of length 1.8. The aspect ratio of the ellipse is 1:3.41, so the cue is classified as not a penny. Cue 5 is resolved to 19 cells, and the best fitting ellipse has a major axis of length 2.57 cells making and a minor axis of length 2.36. The aspect ratio of the ellipse is 1:1.09, so the cue is classified as a penny.

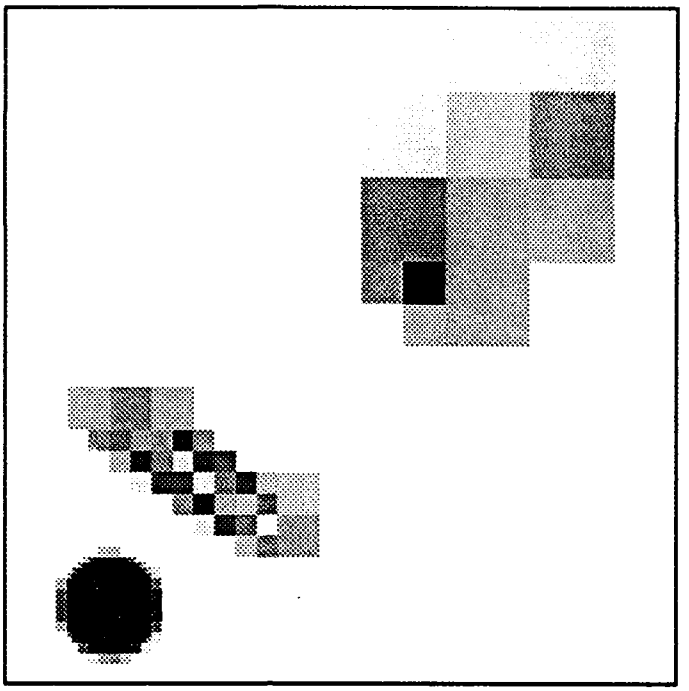


Figure 6.4.1-15. Third foveal sensor frame.

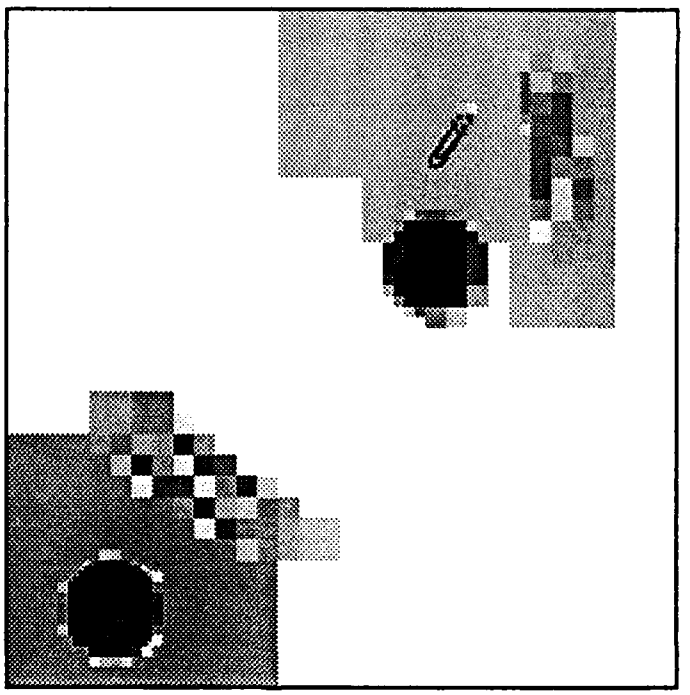


Figure 6.4.1-16. Integrated perception of first three registrations using discard method. The light gray region on the top and dark gray region on the bottom left represent the contribution to the integrated perception by the second and third sensor frames respectively.

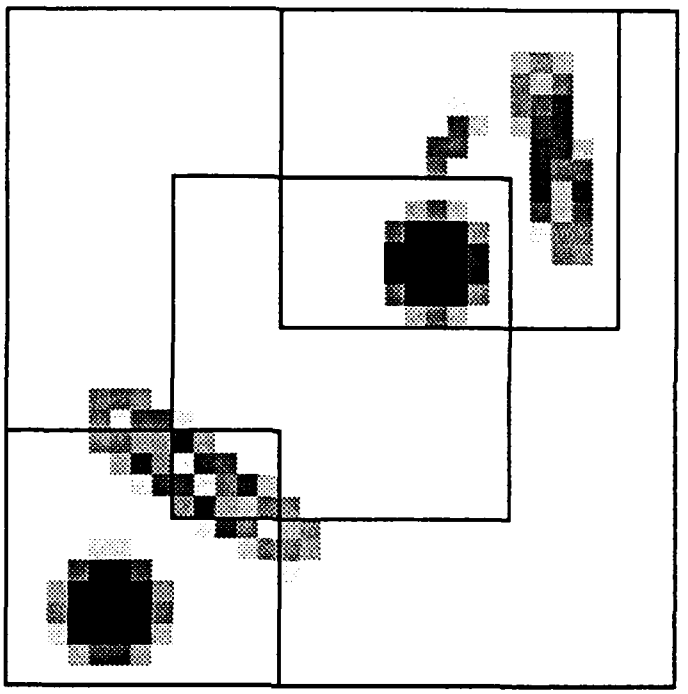


Figure 6.4.1-17. Coverage within a complete pyramid level k=4 by data from the first, second, and third sensor frames.

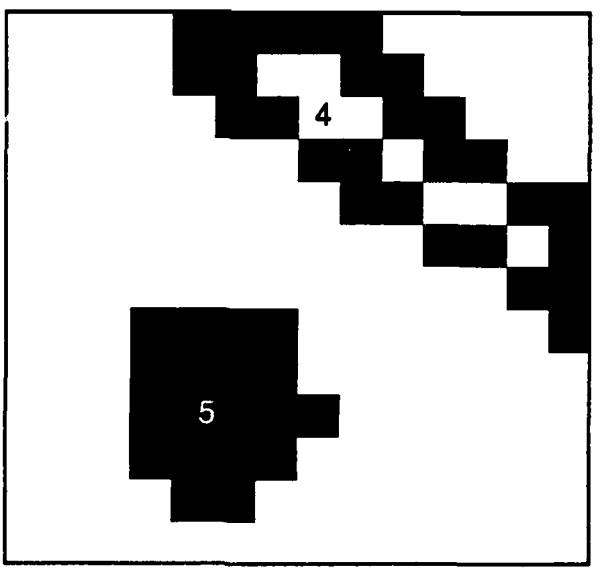


Figure 6.4.1-18. Conditioned and cue labeled polygon level k=4 of third frame. The image is in the coordinates of the region of coverage of the third frame at k=4.

At this point, the top-down analysis has completed processing all the cues, and gives the final result that there are two cents in view. The locations (centroids) of the pennies are known with greater acuity than the initial cues at level k=5 because they are resolved at level k=4. Table 6.4.1-3 summarizes the final results of the vision task. In this exercise, an integrated perception was not necessary: the objects were small enough that the polygons generated from the individual frames were able to represent represent their area with sufficient acuity so as to complete the task.

Cue	Resolution level	Area (in pixels)	Aspect ratio	Label
1	4	368	1:3.94	no penny
2	3	152	1:2.60	no penny
3	4	304	1:1.09	penny
4	4	432	1:3.41	no penny
5	4	304	1:1.09	penny

Table 6.4.1-3. Top down analysis results of pennies exercise.

6.4.2 Identifying the Face of a Card Among Other Objects

In this exercise, the foveal machine vision system is required to resolve with high acuity a small feature of a relatively large object. A model based foveation strategy is employed to align the sensor axis accurately over the small feature. The size difference between the large object and the small feature requires a "homing-in" of the sensor axis from the first perception of the large object to a higher resolution registration of the feature.

The scene, shown in Figure 6.4.2-1, consists of a card, a chip, and a knife. The objective of the machine vision system is to first identify the card from the other objects in the scene, and then identify the face value of the card. A model of a playing card from a poker deck is employed (Figure 6.4.2-2). A *letter region* in which the letter is contained is defined by the model. The objective of the gaze control strategy is to resolve the letter

region of a cue labeled as a card. The letter region is to be resolved to 18 cells or more to support card classification (labeling the card cue as Jack, Queen, or King).

Cues are initially identified at k=5, the lowest level in the polygon in which the entire field-of-regard is supported. The discrimination of the card from the other objects in the scene is performed using an aspect ratio test. A cue with an aspect ratio within 1:1.4±.15 is labeled as a card cue.

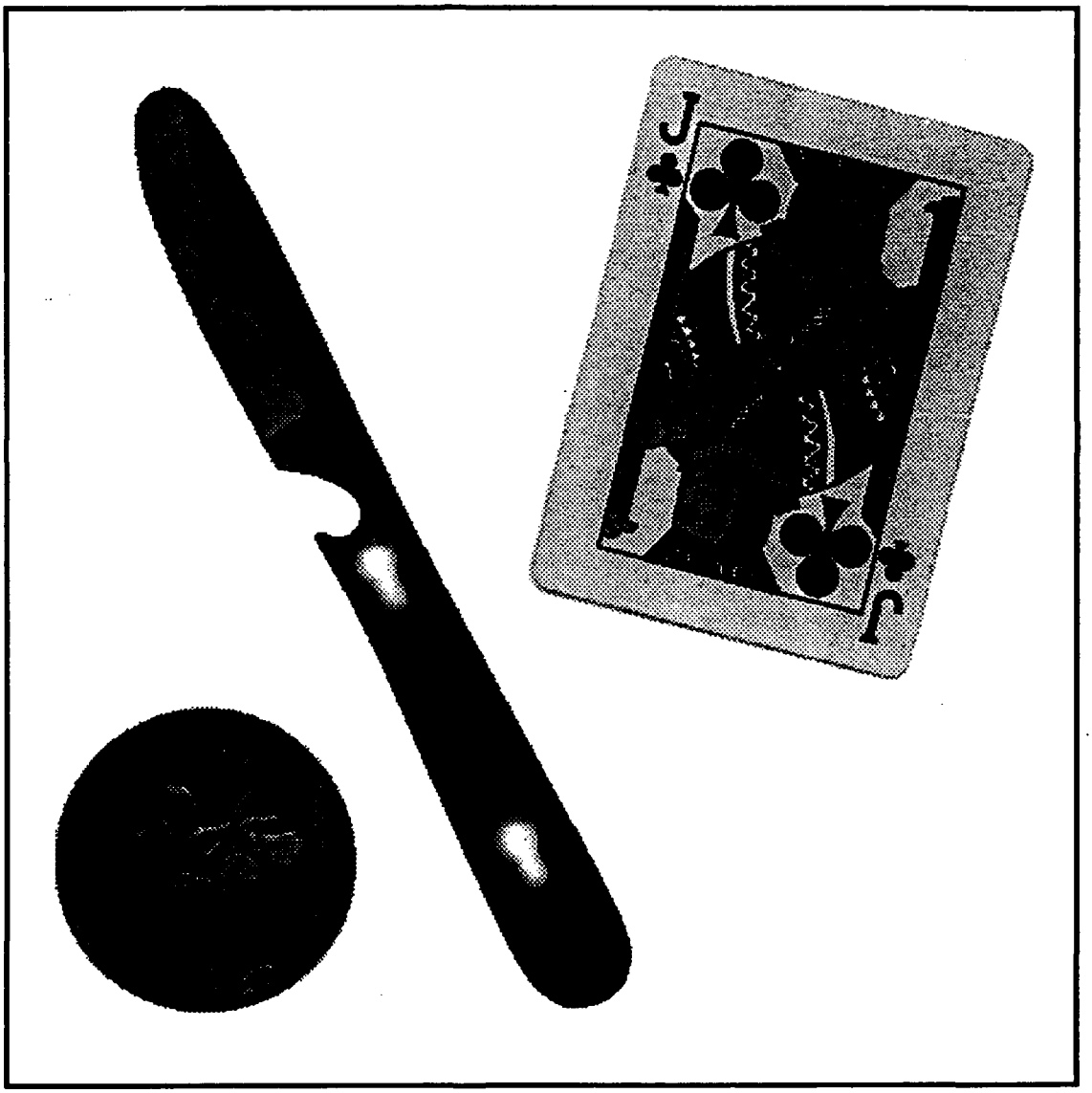


Figure 6.4.2-1. Card scene. The scene consists of 512×512 pixels.

The top-down algorithm for card classification consists of first foveating to the location on the cue labeled as the card where a letter is expected. The features at the corner are then correlated (convolved) with templates for the helvetica letters "J", "Q", and "K", corrected for scale and rotation. The card is classified as the type corresponding to the letter featuring the greatest correlation. The location of the letter is computed by first forming a Laplacian polygon from the Gaussian polygon. The location of the corner of the card is obtained from the lowest level of the Laplacian polygon registering the corner. An offset from the corner to the letter is estimated from the card model, and the area and orientation measurements of the cue. A card cue is resolved to at least 35 cells before making the measurements, to enhance the spatial accuracy of the foveation (at this resolution, the letter region is perceived as an unresolved cue). The Laplacian polygon is employed because it emphasizes edges and can provide better localization of the edge than the Gaussian polygon due to the poor contrast of the card against the background.

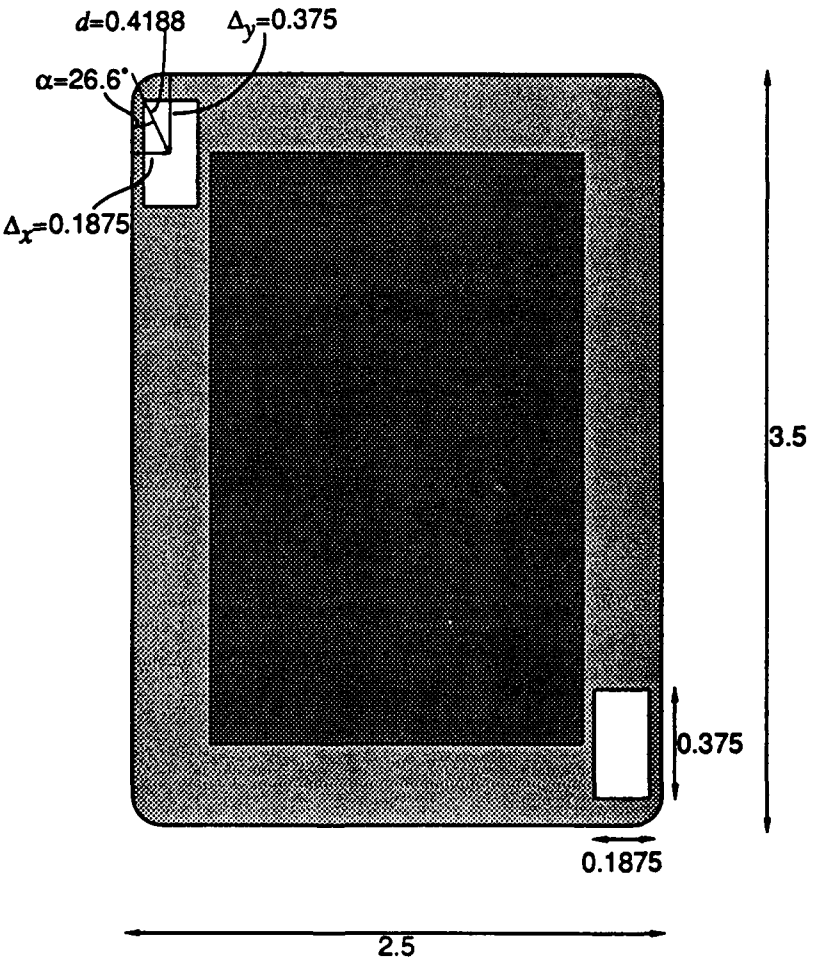


Figure 6.4.2-2. Card model. Units of length are in inches.

The initial registration is made with the optical axis centered over the scene (Figure 6.4.2-3), generating 1216 rexel values. Level k=5 of the polygon, the highest acuity level covering the entire field-of-regard, is shown in Figure 6.4.2-4. As in the previous exercise, the overall polygon is ten levels high (k=0...9).

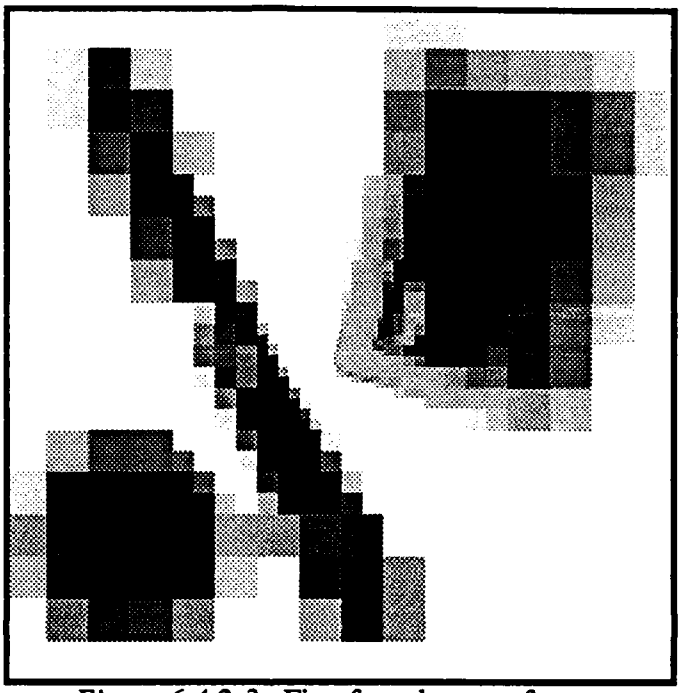


Figure 6.4.2-3. First foveal sensor frame.

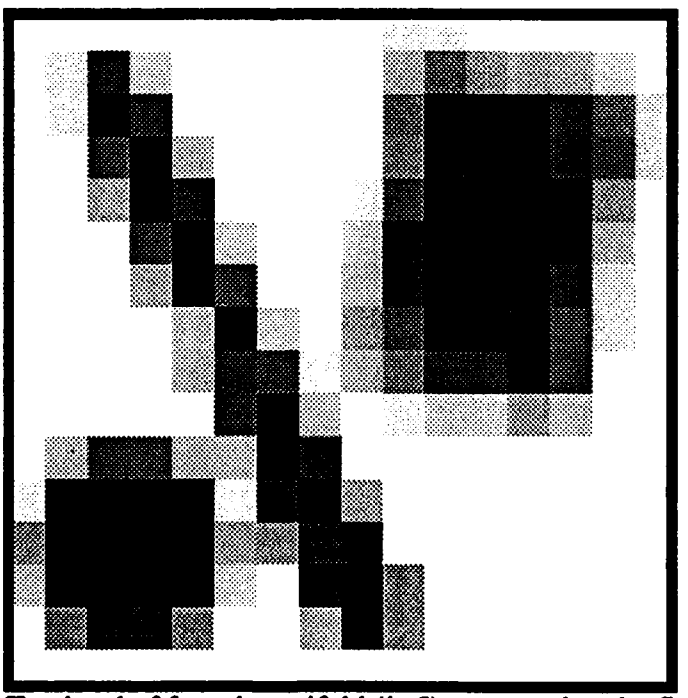


Figure 6.4.2-4. Top level of foveal manifold (k=5) representing the first sensor frame.

After the highpass filtering, histogram equalization, and thresholding of level k=5, three cues are identified (Figure 6.4.2-5) with areas greater than 18 cells. The area analysis results are given in Table 6.4.2-1. Assuming we know *a-priori* that the scene contains a knife and a coin, a cue with an aspect ratio higher than 1:1.4±.15 can be labeled the former, and a cue with lower aspect ratio can be labeled the latter. The aspect ratio of Cue 1 falls exactly in the center of the range for the card, even after loosing some object edge information in the thresholding due to poor contrast between the border of the card and the background. The major axis of the ellipse fitted over cue 1 makes an angle of 76.47° with the horizontal, indicating that the card is tilted 76.47°-90°=-13.53°. The centroid of cue 1 is at (10.5,4.5) in terms of level k=5 coordinates.

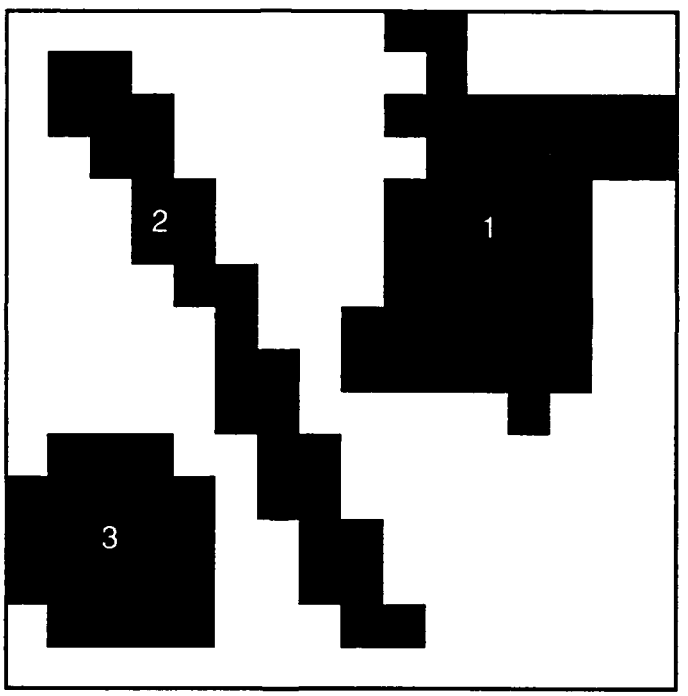


Figure 6.4.2-5. Cue labeling of conditioned and thresholded polygon level k=5.

Cue	Area (k=5 cells)	Aspect ratio	Label
1	44	1:1.4	card
2	28	1:9.1	knife
3	22	1:1.0	coin

Table 6.4.2-1. Top down area analysis results.

Having labeled a cue as a card, the gaze control strategy is now one of foveating directly on the letter region of the cue to resolve the letter on the card. Once this letter region is resolved to 18 cells or more, character recognition is performed. The following steps compute the position of the letter region in the cue.

The ratio of the area of the overall card to that of the letter region is given by the card model as

$$\frac{A_{card}}{A_{letter}} = \frac{2.5 \times 3.5}{0.375 \times 0.1875} = 124.4 \tag{6-23}$$

At level k=4, the letter region of the cue is predicted to be detected as an unresolved dot

$$k_{detect} = 5 - \left\lceil \log_4\left(\frac{125}{44}\right) \right\rceil = 5 - \lceil 0.753 \rceil = 4 \tag{6-24}$$

and at level k=2, the letter region is predicted to be resolved to the *a posteriori* determined acuity

$$k_{resolve} = 5 - \left\lceil \log_4\left(\frac{18}{\frac{44}{125}}\right) \right\rceil = 5 - \lceil 2.84 \rceil = 2 \tag{6-25}$$

At level k=2, this region is expected to be resolved to approximately 23 cells

$$\#\ \text{of cells} = 4^{5-2} \times \frac{44}{125} \cong 23 \tag{6-26}$$

The top left hand corner of the card is determined from a Laplacian polygon, and the system foveates to an offset of this location where the letter is expected. The offset is computed from the relative geometry of the model quantified by the measured scale and orientation of the card in the scene. The linear scale u_k of the card at level k is computed from the area measurement by

$$u_k = \sqrt{\frac{A_k}{2.5 \times 3.5}} = \sqrt{\frac{A_k}{8.75}} \tag{6-27}$$

where A_k is the cue area measured at level k. Given the area measurement of 44 cells at level k=5, u_5 is 2.24 cells. The scales at different levels are related by

$$u_{k_2} = 2^{k_1 - k_2} u_{k_1} \tag{6-28}$$

The offset in level k of the letter from the left hand corner of the card is

$$d_k = 0.4188u_k \tag{6-29}$$

$$\beta = \theta - 63.4^\circ \tag{6-30}$$

where θ is the orientation of the cue (specifically the angle between the major semiaxis of the best fitting ellipse in k=5), measured to be -13.53°. In Cartesian terms, the offset is

$$d_{k,x} = d_k \cos\beta = 0.0947u_k \tag{6-31}$$

$$d_{k,y} = d_k \sin\beta = 0.40795u_k \tag{6-32}$$

After foveating to the letter region, the distance between the optical axis and the corner of the cue is measured. When this distance is measured at level $k_{resolve}$+1 to within 17% of the model based distance

$$d_{3,x} = 0.0947u_3 = 0.0947 \times 2^{5-3}u_5 = 0.849 \tag{6-33}$$

$$d_{3,y} = 0.40795 \times 2^{5-3}u_5 = 3.655 \tag{6-34}$$

then the letter region is contained within polygon level $k_{resolve}$ and template matching can be performed. If the distance is greater, then the foveation process is repeated using the new higher resolution registration of the top left corner to further home-in on the feature.

The location of the corner is now obtained. Because level k=5 is the lowest level in the Gaussian polygon registering the top left corner of the card cue, level k=5 will also be the lowest level in the Laplacian polygon registering the corner. Consequently, only this level of the Laplacian polygon is necessary. The Laplacian polygon level k=5 is obtained by taking the difference between Gaussian polygon level k=5 and the expanded subset of level k=6 with the same spatial coverage, and is given in Figure 6.4.2-6. The dynamic range of the features are expanded through histogram equalization (Figure 6.4.2-7). We are interested in light to dark transitions so the dark features of the equalized Laplacian level will be measured.

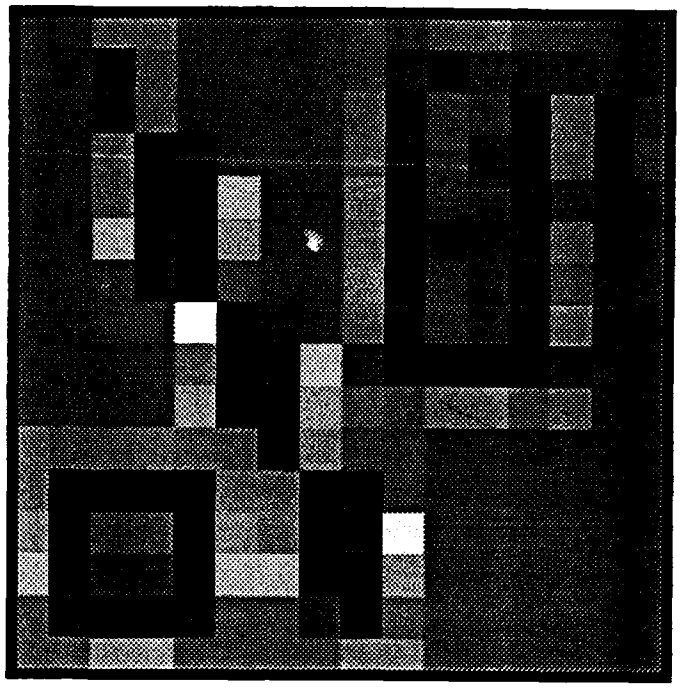


Figure 6.4.2-6. Laplacian polygon level k=5 of first sensor frame.

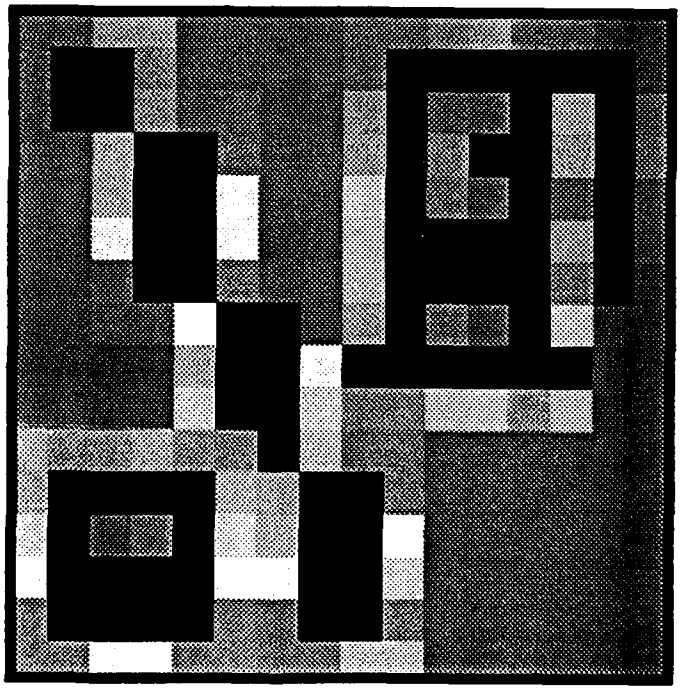


Figure 6.4.2-7. Histogram equalized Laplacian level k=5 of first sensor frame.

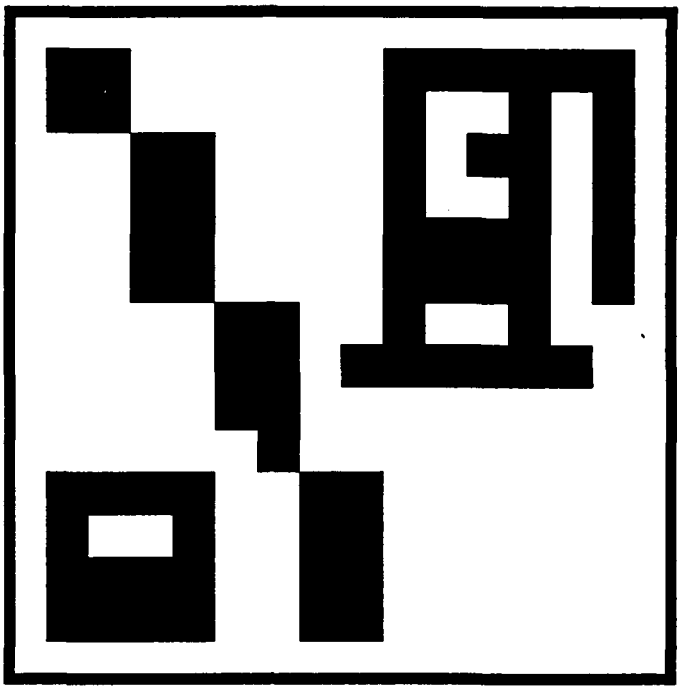


Figure 6.4.2-8. Equalized and thresholded Laplacian level k=5 of first sensor frame.

The top left cell of the thresholded Laplacian level k=5 (Figure 6.4.2-8) is found to be at (9,2) in level k=5 coordinates. The center of this cell in level k=0 coordinates is (304,80). Note that conditioning of the Laplacian polygon levels does not include explicit edge enhancement (high pass filtering); this is automatically performed by the differencing operation of the bottom-up generation process forming the Laplacian polygon.

The second sensor frame is given in Figure 6.4.2-9. The top left corner of the card cue is resolved at level k=3 (Figure 6.4.2-10) and measured in the Laplacian level k=3 (Figure 6.4.2-11,12) to be of vector (-1,7) cells from the axis, which is outside the range necessary to register the letter region with polygon level k=2. Indeed, as is seen in Figure 6.4.2-13, the sensor axis is too low to register the entire letter at level k=3, although part of it is perceived. This error in foveation originates from the measurement of the location of the cue top left corner at level k=5 of the first Laplacian polygon. In the second polygon, the corner is resolved at k=3 and the true location is better measured. The foveal axis location is moved 1+0.849 cells to the right and 7-3.655 cells upward, in k=3 coordinates, or 15 pixels to the right and 27 pixels upward in k=0 coordinates.

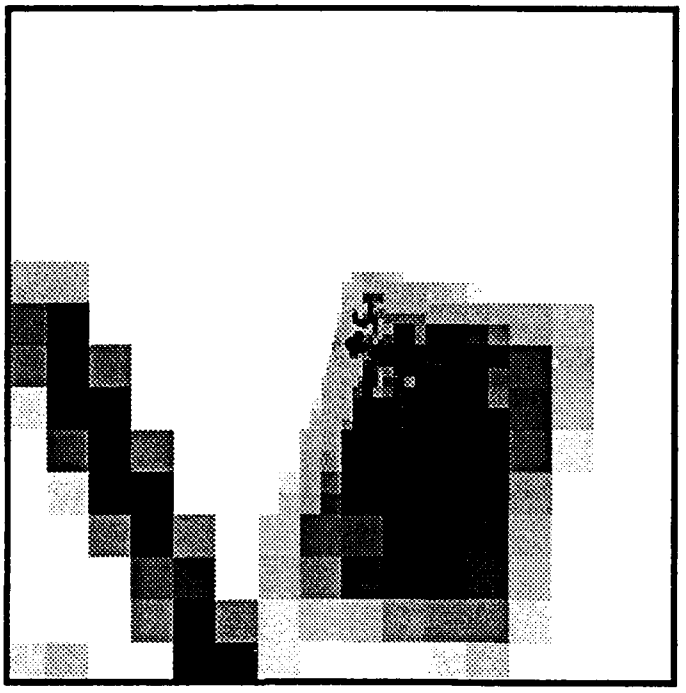


Figure 6.4.2-9. Second foveal sensor frame.

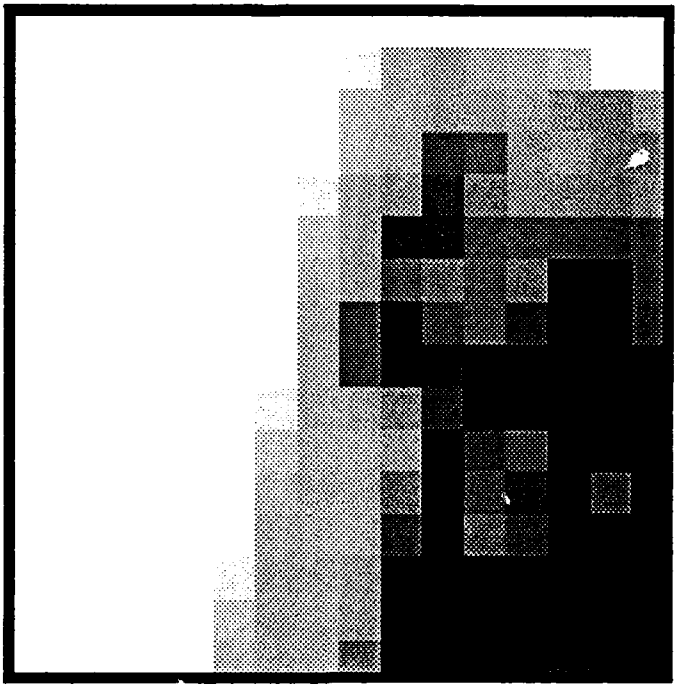


Figure 6.4.2-10. Gaussian polygon level k=3 of the second sensor frame.

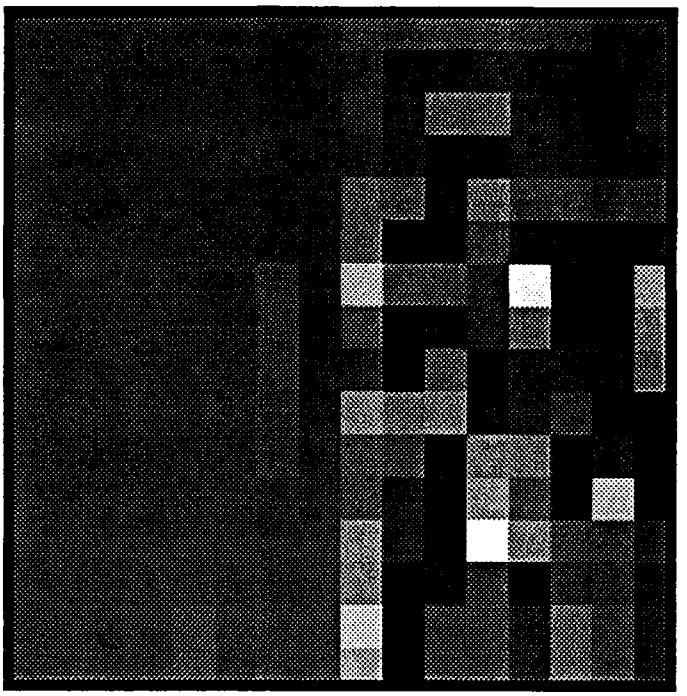


Figure 6.4.2-11. Laplacian polygon level k=3 of second sensor frame.

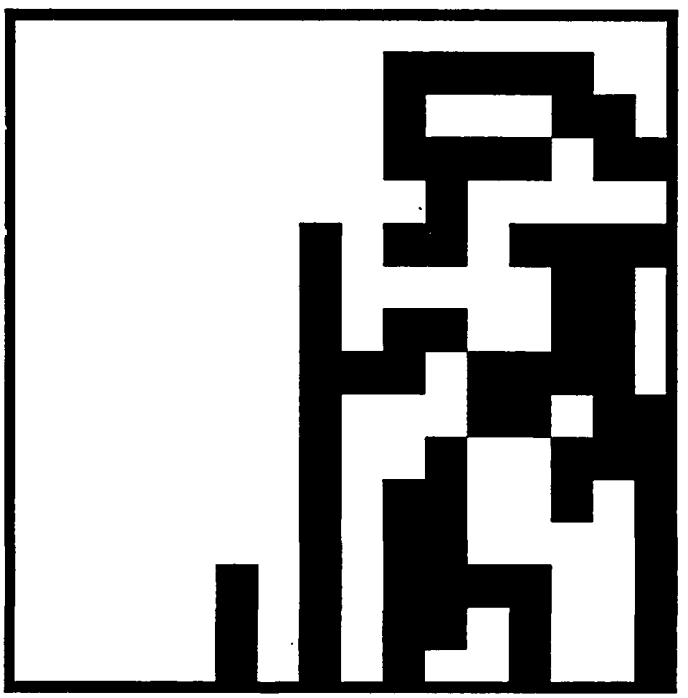


Figure 6.4.2-12. Equalized and thresholded Laplacian level k=3 of second frame.

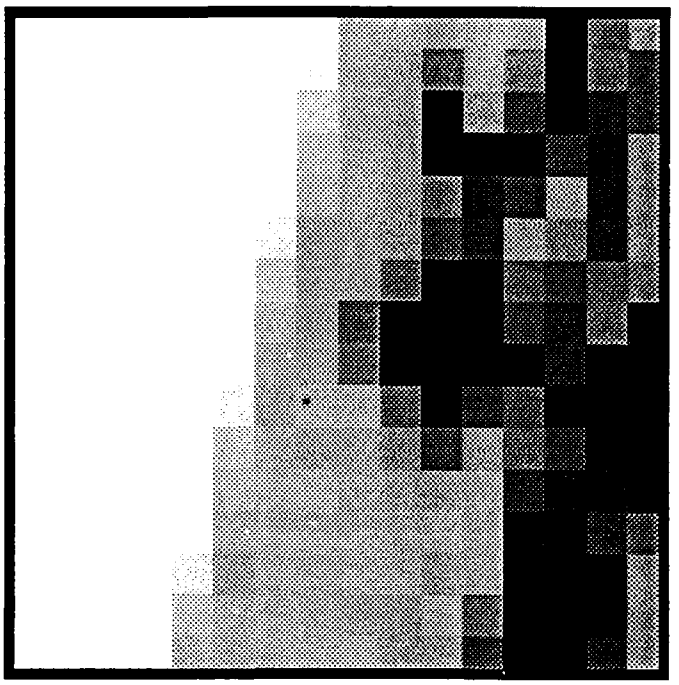


Figure 6.4.2-13. Gaussian polygon level k=2 of the second sensor frame.

The third sensor frame is given in Figure 6.4.2-14. Level k=3 of the Laplacian polygon (Figure 6.4.2-15) confirms that at this acuity, the sensor axis is centered over the letter region of the card cue, so template matching can now be performed. This operation is performed at level $k_{resolve}$=2 of the Gaussian polygon (Figure 6.4.2-16). The level is first convolved with an edge detection kernel to remove any shading (Figures 6.4.2-17,18) because the letter templates are composed of line and curve segments. The Laplacian polygon is not used for template matching because the difference of Gaussians filtering is not as good an edge detector.

The master templates (Figure 6.4.2-19) are 18 pixels tall. Before using these templates as convolution kernels, they must be scaled and rotated as the letter on the card. From the area analysis of the first polygon, the rotation is known to be -13.53° and the scale factor at k=2 to be u_2=17.92 cells. The model letter is 0.375 units tall, so the unrotated templates must be approximately seven (17.92×0.375=6.72) cells tall at level k=2. The master templates are rotated by -13.53° and scaled by a factor of $\frac{6.72}{18}$ =0.37 through an interpolative process (Figure 6.4.2-20) and thresholded to give the actual convolution kernels. The peak values in the convolution results are 204 (out of a maximum of 255) for "J", 152 for "Q", and 129 for "K", with the higher values indicating greater correlation. The card is thus determined to be a jack. The suit could be determined by

following the card model and comparing the feature against templates for heart, diamond, club, and spade. If the templates require greater resolution to discriminate between the features, the system will have to foveate a little lower.

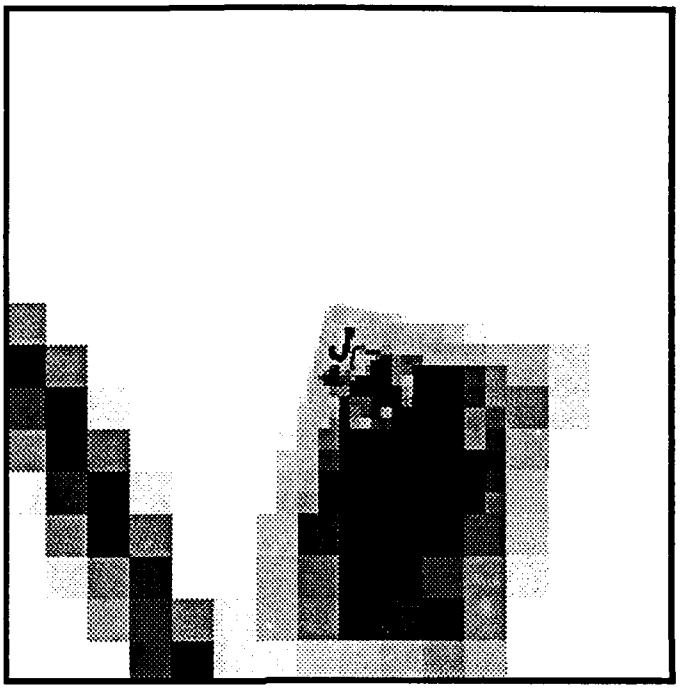


Figure 6.4.2-14. Third foveal sensor frame.

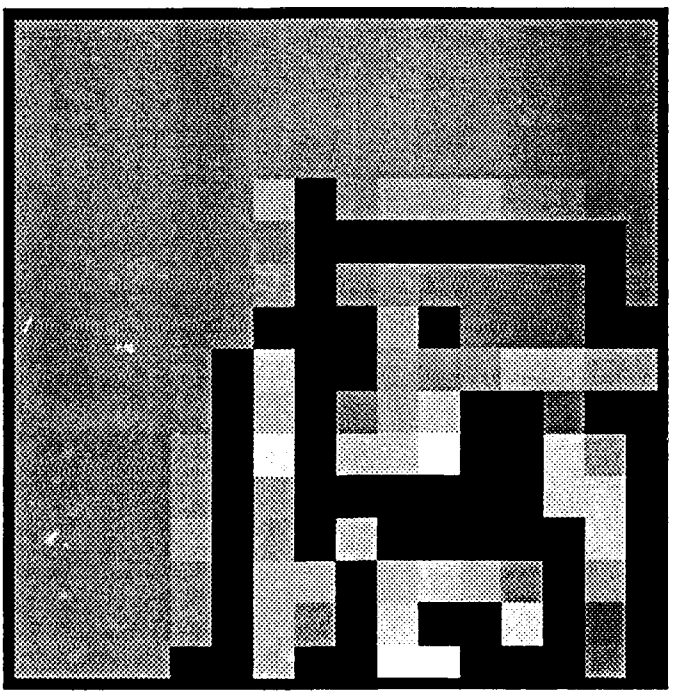


Figure 6.4.2-15. Equalized Laplacian level k=3 of third frame.

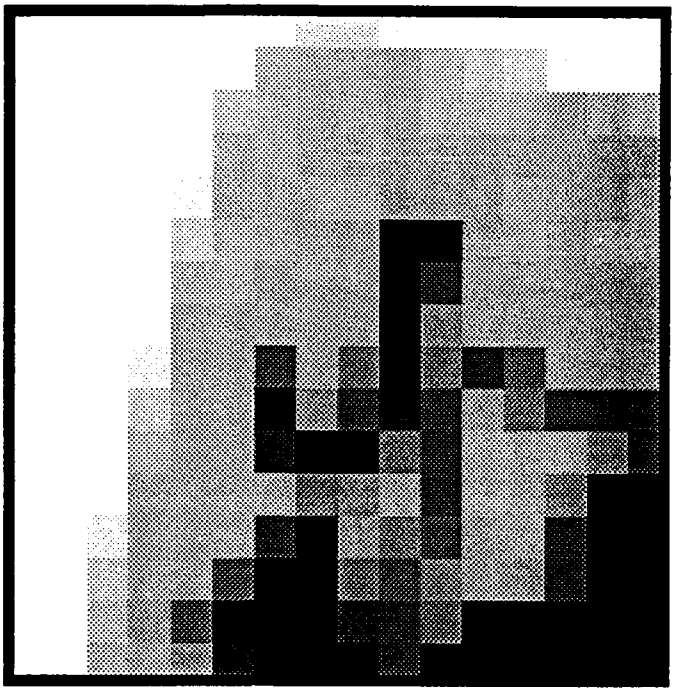


Figure 6.4.2-16. Gaussian polygon level k=2 of third frame.

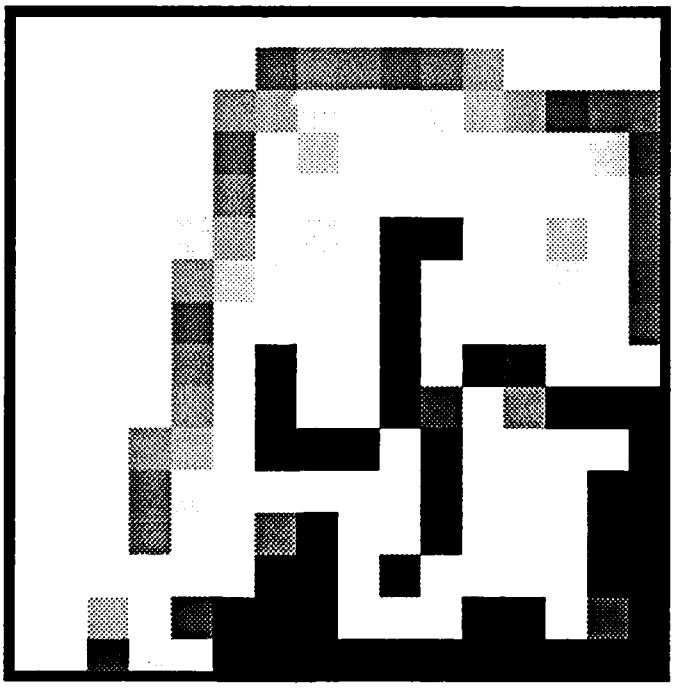


Figure 6.4.2-17. Edge detection filtered Gaussian polygon level k=2 of third frame.

-1	-1	-1
-1	8	-1
-1	-1	-1

Figure 6.4.2-18. Edge detection filter kernel.

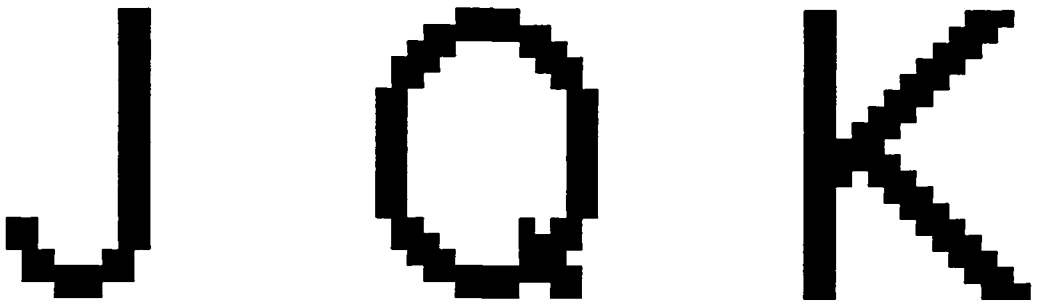


Figure 6.4.2-19. Master letter templates.

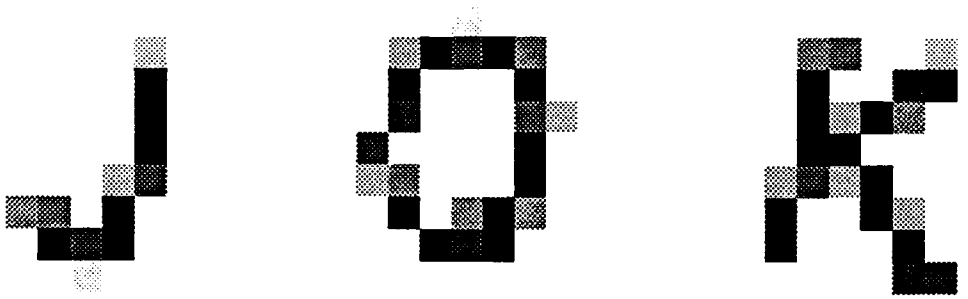


Figure 6.4.2-20. Scaled and rotated letter convolution kernels.

The homing-in process performed so well that the letter region is covered (just barely due to scaling) by the relatively small k=1 level of the third frame (Figure 6.4.2-21). If the letter region had to be resolved to more than 18 cells, the third frame would have satisfied this requirement up to approximately 4×23=92 cells. If the card scaling was larger, then level k=2 could be employed.

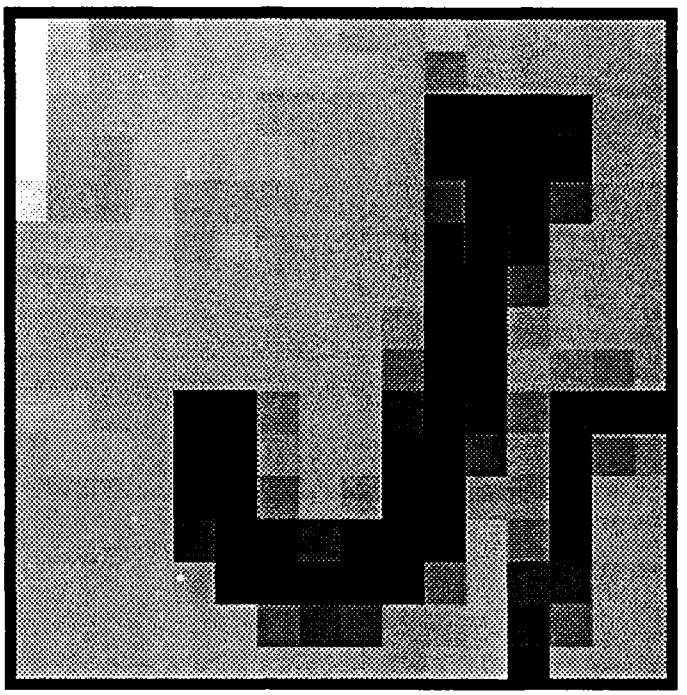


Figure 6.4.2-21. Gaussian polygon level k=1 of third frame.

6.4.3 Additional Exercises

This section summarizes a set of eight experiments conducted with different scenes (Figure 6.4.3-1). The task of the foveal system was to identify the jack, queen, and king playing cards in the scene. The algorithm employed (Figure 6.4.3-2) is similar to that used in the preceding exercises; cell value thresholding is used to identify objects, and the top left corner of any object with as aspect ratio near 1:1.4 is interrogated for the presence of an upper case letter "J", "Q", or "K".

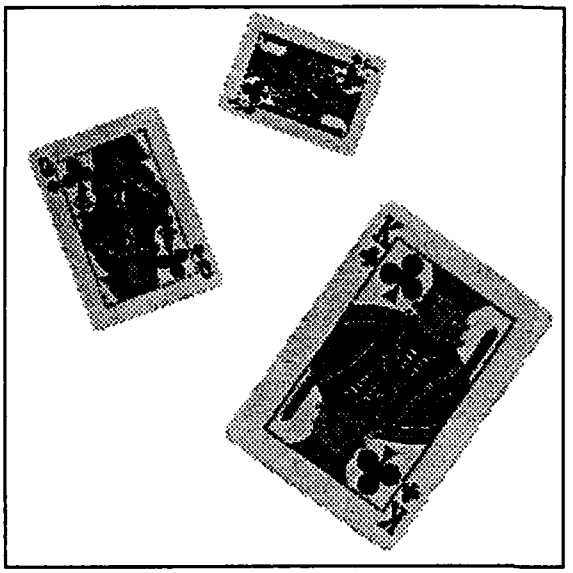
An assumption made by the algorithm is that the background is the predominant feature in the initial field-of-regard. This implies that there are more cells at the waist of the Gaussian polygon generated from the first frame with values derived from averaging background pixel values than with values derived from object pixels. Another assumption made by the algorithm is that the grey level distribution of the background is unimodal at the polygon waist. The algorithm employed in this section distinguishes background cells from object cells by computing the histogram of cell values at the waist, identifying the

main lobe in the histogram, and labeling all cells with values in the main lobe as background cells.

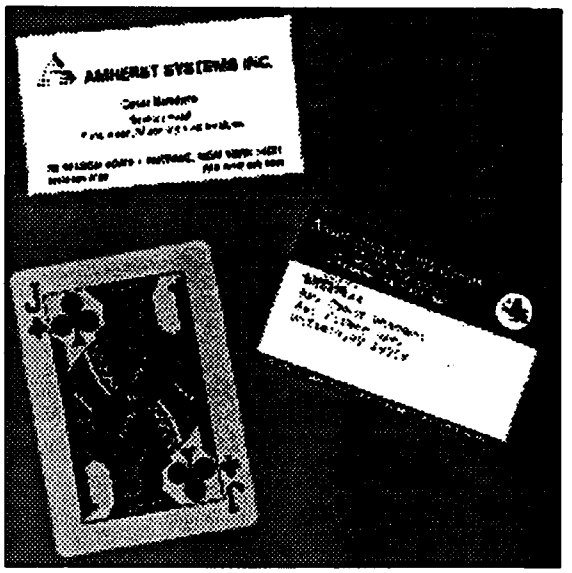
The algorithm employed in this section does not perform temporal filtering. However, when the main histogram lobe is wide (40% of the eight bit dynamic range), and letter interrogation fails, the algorithm labels the interrogated object as undetermined. This reduces the probability of false negative classification in scenes with wide clutter or noise variance, because template matching discrimination suffers in such conditions. Also, the main lobe can include the values of object cells, causing the labeling of object cells as background cells and erroneous area, aspect ratio, and orientation measurements.

The results of the eight exercises are summarized in Table 6.4.3-1. Only a small number of foveations were executed in each exercise, and each foveation generated a polygon of 1.6% the size of a corresponding pyramid. There were no false positive or false negative classifications. In the first exercise, the system had to foveate to the center of the smaller card cues to obtain the area and orientation measurements with sufficient resolution. In the second exercise, the jack and the rightmost card passed the aspect ratio test, but the latter was rejected by the letter test. In the third exercise, both objects failed the aspect ratio test. In the fourth exercise, all non-card objects failed the aspect ratio test. In the fifth and sixth exercises, the coin and knife failed the aspect ratio test, and the jack passed the letter test. Even though the noise is bimodal, it is averaged to a narrow unimodal distribution at the polygon waist.

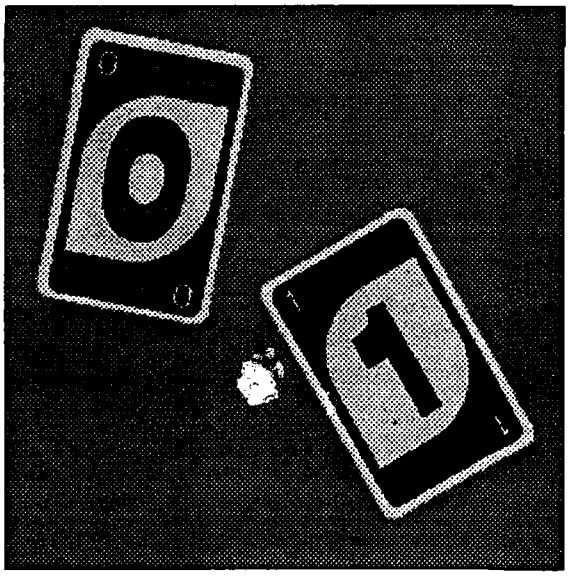
In the seventh exercise, the noise variance was sufficiently wide so as to cause some card cells to be erroneously labeled as background cells. This resulted in an object orientation measurement error of 14°. The mismatch between template and letter orientation, plus the extreme noise, resulted in poor template matching discrimination. The coin object passed the initial aspect ratio test. However, upon foveating to the cue to obtain higher resolution measurements, its aspect ratio was measured to be below the range for cards, and the cue was dropped. In the eighth exercise, the paragraph text and the blank space between the paragraphs produced a bimodal distribution at the polygon waist, violating an algorithm assumption. Consequently, when labeling the predominant lobe (corresponding to the text) as background, the spacing between the paragraphs was labeled as an object which connected cues and corrupted the area and aspect ratio measurements.



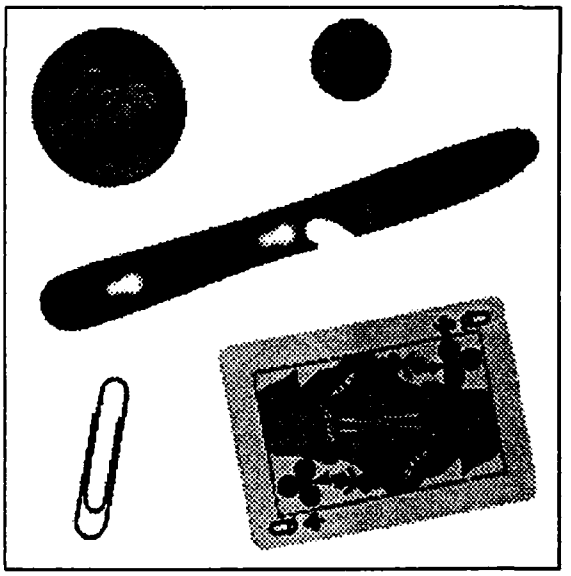
Scene 1. Jack (50% relative scale), queen (66.6%), king (100%).



Scene 2. Business card, society membership card, jack.

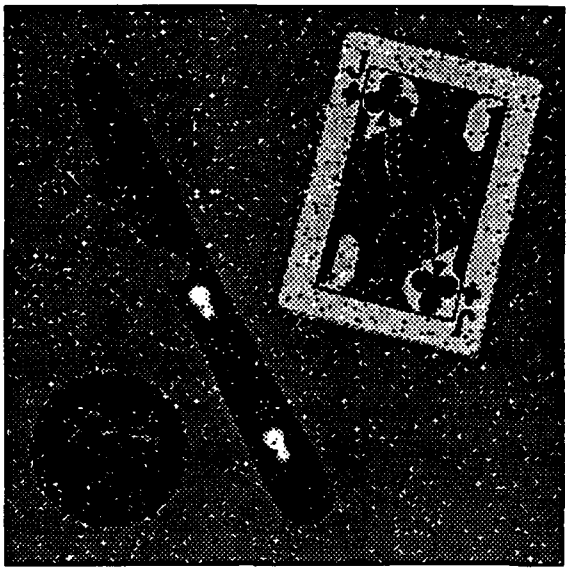


Scene 3. Non-poker playing cards.

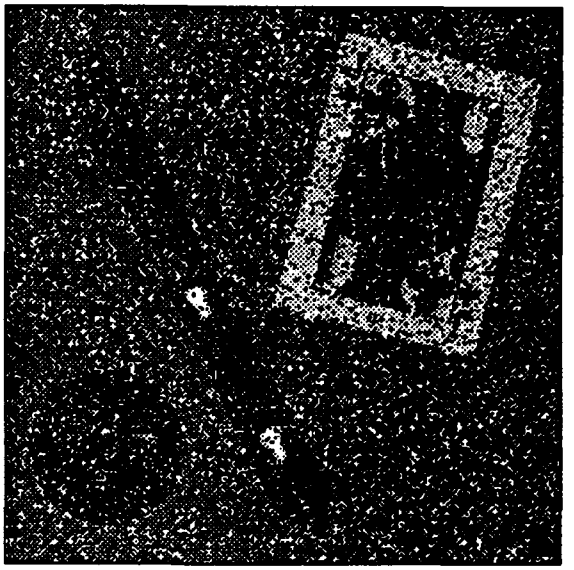


Scene 4. Queen among many other objects

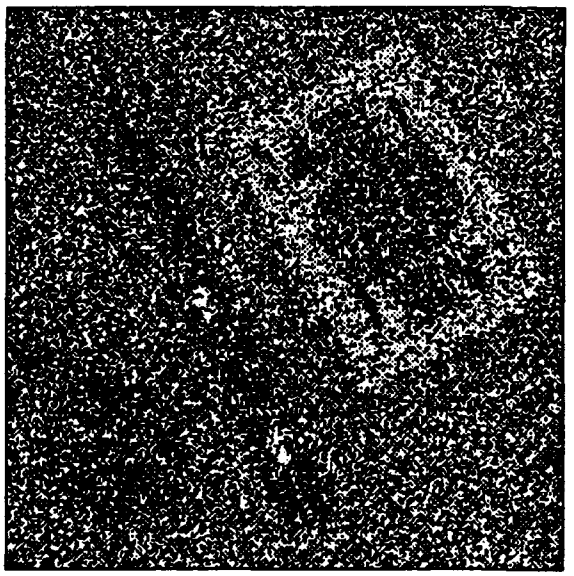
Figure 6.4.3-1. Exercise scenes. (continued into next page)



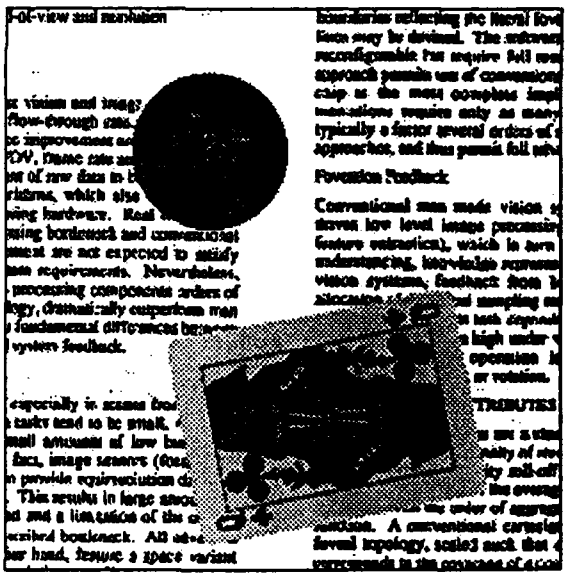
Scene 5. Jack and other objects with 5% of pixels set to black or white.



Scene 6. Jack and other objects with 22% of pixels set to black or white.



Scene 7. King and other objects with 50% of pixels set to black or white.



Scene 8. Queen and coin over bimodal background.

Figure 6.4.3-1. Exercise scenes. (continued from previous page)

```
Foveate to center of field-of-regard; Generate Gaussian polygon;
Group cells into objects at waist of polygon:
    Compute cell value histogram of waist;
    Identify main lobe of distribution;
    If (width of main lobe is greater than 40% of dynamic range) then label scene as noisy; else
    label scene as free from significant noise;
    Label waist cells with values within main lobe as background cells; label remaining cells as object
    cells;
    Label each convex hull of object cells as a unique object;
    Label touching convex hulls as separate objects;
Measure area of objects at waist of polygon;
Disregard objects smaller than four cells;
Measure centroid, aspect ratio, and major axis orientation of remaining objects;
For (every object with an aspect ratio within 1:1.4±0.15) do {
    If (the object has an area < 35 cells) then {
        Compute the minimum resolution which will resolve the object to ≥ 35 cells;
        If (object not in view at minimum resolution) then {Foveate to object centroid; Generate
        Gaussian polygon; };
        Repeat area and orientation measurement at lowest polygon level with object in view; };
    Compute scale and rotation of letter region using playing card model and object area, orientation
    measurements;
    Compute the minimum resolution which will resolve the letter region to ≥ 35 cells;
    do { Locate top left object corner at lowest polygon level with corner in view;
        Compute center of letter region using playing card model;
        If (letter not in view at minimum resolution) then {Foveate to letter region; Generate Gaussian
        polygon};          }
    while (letter not in view at minimum resolution);
    Rotate and scale letter templates to that of letter region at lowest polygon level k with region in view;
    Convolve letter templates with polygon level k; Identify peak value of each convolution within center
    50% of letter region;
    If (all peaks above 150 out of maximum correlation vale of 255 ) and (convolution peak for one letter is
    33% closer to 255 than the other peaks) then
        label object as a playing card with that letter;
    else {
        If (scene free from significant noise) then
            label object as not a jack, queen, or king ;
        else
            label object as undetermined;    };        };
```

Figure 6.4.3-2. Pseudocode of foveal system algorithm.

Scene	Total number of objects in scene	Number and classification of objects labeled as cards / actual poker cards	Total number of foveations
1	3	3 (J, Q, K) / 3 (J, Q, K)	6
2	3	1 (J) / 1 (J)	3
3	2	0 / 0	3
4	5	1 (Q) / 1 (Q)	2
5	3	1 (J) / 1 (J)	2
6	3	1 (J) / 1 (J)	2
7	3	1 (?) / 1 (K)	3
8	2	1 (?) / 1 (Q)	1

Table 6.4.3-1. Top down analysis results of card exercises.

6.4.4 General Conclusions from Exercises

The preceding exercises illustrate several strengths and weaknesses of foveal vision. Weaknesses in the algorithms employed should not be misconstrued as those of the concept of foveal vision. On the other hand, a machine system is only as good as the algorithms employed, and the graded resolution of foveal vision (i.e., the fact that polygon levels below the waist do not cover the entire field-of-view) affects algorithm performance. Consequently, discussions on the performance of foveal vision should be in the context of the acuity profile and the vision and gaze control algorithms.

The exercises illustrate the significance of an object model with efficiently described features (e.g., features represented at a resolution which neither oversamples nor undersamples the relevant properties of the feature) spatially related so as to form a feature graph or feature tree. Such models are critical to the process of scene understanding in machine vision, where objects and their states (e.g., position, attitude, angles at articulations) are identified by following the branches of the graph in a top-down fashion and interrogating image space to measure properties and confirm or deny the presence of features. In foveal vision, the model has the additional role of directing the sensor fovea.

One problem which can occur in foveal image processing is that a polygon level with the necessary resolution to resolve a feature may not be able to register the entire feature. A feature interrogation algorithm which operates only at a single level would have to process only part of the feature, process a higher level (lower resolution) feature representation, or both. This can lead to suboptimal performance in algorithms designed to operate on the entire feature, such as template matching.

One way to overcome the spatial extent limitations of the lower polygon levels is by forming and processing a polygon from a low level integrated perception, thus extending the effective coverage of the polygon levels beyond that of a single sensor frame. This approach, however, is not appropriate when the feature is moving with respect to the sensor because low level frame-to-frame correlation becomes difficult or impossible.

Another approach is to use algorithms which can operate on a feature with different portions represented at different resolutions. This may require of the object model the representation of features in a multiresolution format, such as the representation of the object itself. Alternatively, two or more polygon levels may be collapsed into a uniform array by replicating (expanding) the cells of the outer 75% of the higher level not supported at the lower level. The images of foveal sensor frames in this dissertation were generated using this planar representation of multiresolution data. This approach is inefficient with respect to data storage because data is replicated. However, the overhead of expanding a small number of levels can be more than compensated by the computational tractability of a uniform data array.

A third approach is to ensure that the features in the object model are defined such that the number of cells representing the feature never exceeds the number of cells of the polygon level expected to be employed in the interrogation. The expected level for features corresponding to object parts, such as the letter of a playing card (as opposed to features of the overall object such as the card aspect ratio), will often be below the waist. In these cases, the representation limits are fixed to $4d \times 4d$, where d is the foveal pattern subdivision factor (assuming an exponential geometry), and the approach is scale invariant.

There are several strong reasons for the decomposition of objects into parts or features [Hoffm86]. First, the vision system is less sensitive to the occlusion of object portions because it can identify the visible objects. Also, the feature tree is typically a much more compact data structure than the integral representation of the object at all possible

aspect angles. Second, object articulations are naturally handled by decomposition whereas they increase the dimensionality of the integral representations. A third reason of particular importance to foveal vision is that the resolution of the model should be localized.

All the above approaches to overcome the spatial extent limitations of higher acuity in foveal vision benefit from a smaller feature bandwidth × feature size product. In the case of forming an integrated perception, fewer frames need to be integrated, increasing temporal resolution and reducing data storage requirements. In the case of level collapsing, fewer levels need to be expanded, permitting more of the feature to be registered at the higher resolution and also reducing data storage requirements. Of course, the design of an object model with localized feature bandwidth and size is also simplified when the object lends itself to decomposition.

Just as the interrogation of object features should operate at multiple resolutions to better employ the polygon data, so should object detection. For example, if detection is limited to a single level, then the discrimination of objects from background may interpret closely spaced objects as one. One approach with pyramid data is to repeat this labeling process on object areas at lower levels in search for background features which might segment a hypothesized object into an object cluster. This approach, being a top-down algorithm, can be executed on a foveal polygon and drive the gaze control strategy.

6.5 Some Comparisons between Foveal and Pyramid Processing

The relative benefits and limitations of foveal polygon processing with respect to conventional pyramid processing stem from the fact that the polygon data structure is a subset of the pyramid. One obvious benefit is the polygon's smaller size. From (3-13) and (3-92), the size of a uniform resolution frame with the same field-of-view and maximum resolution as an exponential pattern with r major rings subdivided by a factor d is

$$A_p = d2^{r+1} \times d2^{r+1} \tag{6-35}$$

The size of the pyramid hierarchical data structure generated from such a frame is given by (6-7) as

$$D_{pyra} = \frac{4}{3}d^2 2^{2r+2} - \frac{1}{3} \cong \frac{d^2}{3}2^{2r+4} \tag{6-36}$$

The size of the foveal polygon hierarchical data structure generated from an exponential frame is given by (6-15) as

$$D_{poly} = 16d^2\left(r + \frac{1}{3}\right) \tag{6-37}$$

and the size ratio of the two data structures is

$$\frac{D_{pyra}}{D_{poly}} = \frac{2^{2r}}{3r+1} \tag{6-38}$$

Note that the ratio is not a function of the subdivision factor. As d is increased, the field-of-view (or resolution, depending on final geometry normalization) of both data structures is increased uniformly. Just as the difference in data size between foveal and uniform acuity frames is significant, so is that of the resulting hierarchical databases (Table 6.5-1). These savings can result in significant computational savings when the number of foveations is small.

Number of major rings r	Field-of-view, d=1	Field-of-view, d=4	Field-of-view, d=8	$\frac{D_{pyra}}{D_{poly}}$
4	32×32	128×128	256×256	19.7
6	128×128	512×512	1024×1024	215.6
8	512×512	2048×2048	4096×4096	2,621
10	2048×2048	8192×8192	16K×16K	33,825
12	8192×8192	32K×32K	64K×64K	453,438

Table 6.5-1. Sizes of foveal polygons from single centered frames. The foveal geometry is an exponential pattern with r major rings uniformly subdivided by a factor d.

In the card exercise of Section 6.4.2, a foveal system with six rings and a subdivision factor of four generated three single frame Gaussian polygons prior to completing its task, each consisting of 1878 values. The polygons were processed sequentially, and only the polygon for the most recent frame had to be retained, so the total storage requirements remained at 1878 scalar values. A total of 5634 data values were generated by or derived from the three foveal frames. The uniresolution alternative would have been a pyramid from a single frame with the same field-of-view and maximum resolution as the foveal sensor (512×512 pixels), occupying approximately 350,000 data values.

Parallel processing architectures can be built which process pyramid data structures much faster than a uniprocessor. Such an architecture is called a *pyramid computer*. One implementation of this architecture employs a sequence of 2-D processor arrays hierarchically connected such that each processor in one array is connected to its corresponding sibling processors in the adjacent (lower) array. A processor volume is thus formed which resembles the data structure, and the processing of a data structure cell is assigned to the corresponding processor [Schae87], [Tanim87]. The largest processor array is of course the one representing level k=0, containing one processor for every pixel in the field-of-view, whereas the smallest array is the apex level with a single processor. A drawback of such pyramid computers is the extensive number of processors required.

Parallel processing architectures can also be built for polygon data structures. Such an architecture is called a *polygon computer*. Like the pyramid computer, the polygon computer may be implemented by connected processor arrays so as to form a processor volume which resembles the data structure. The size of an array in this polygon computer is $2^{R-k} \times 2^{R-k}$ from the apex (k=R) to the waist (k=R-$\log_2 d$-2), as in the multiprocessor computer for a pyramid data structure with the same field-of-view ($2^R \times 2^R$) and maximum resolution. The levels at and below the waist are of size $4d \times 4d$ forming a piped architecture, in contrast to the pyramid arrays which keep on growing.

The number of processors in a polygon computer is much less than the number of processors in a pyramid multiprocessor computer. The difference is equal to that between the data structure sizes. For example, a pyramid computer processing frames of size 512×512 has two orders of magnitude more processors than a polygon computer processing exponential patterns with eight rings and a subdivision factor of four, as in the previous exercises.

Another significant difference between the polygon and pyramid computers is that at the levels from the base to the waist, 75% of the polygon processors have no siblings. This greatly simplifies the processor interconnection scheme of the polygon. Only 25% of the processors, forming the array center, require the nine-way processor interconnection scheme of the pyramid (four nearest coplanar neighbors, four siblings, one parent). The processors at the four corners of the array require only a three-way scheme (two neighbors and a sibling), the $12d$-4 edge processors require a four way scheme (three neighbors and a sibling), and the remaining require a five way scheme (four neighbors and a sibling). The processor arrays at these levels, which form the majority of the polygon, are identical; their size and 3-D interconnectivity are invariant. The reduced number of processors and interconnectivity can make polygon computers more feasible to manufacture. Furthermore, interconnection invariance can make polygons computers easily implementable using other parallel architectures [Mares88].

Implementations of pyramid computers typically employ a single processor array large enough to support the base level [Burt88], [Canton86], [Canton88], [Stout88]. The different levels of the pyramid data structure are processed by this array. The advantage of this implementation is the savings in processors (the base level accounts for 75% of all pyramid cells), and that a simple 2-D four-way nearest neighbor processor interconnection scheme can be used, as opposed to the much more complex 3-D nine-way scheme.

A polygon computer may also be implemented with a single processor array. The polygon processor array size is only $4d \times 4d$, as opposed to the much larger pyramid processor array size of $d2^R \times d2^R$ where d is the subdivision factor, and $d2^R$ is the linear dimension of the field-of-view.

In applications where a low level integrated perception is not formed and each foveation generates a polygon data structure, the database size ratio indicates how many additional foveations are allowed before the foveal system manipulates more data than a uniresolution system with the same field-of-view and maximum resolution. It is seen from Table 6.5-1 that many foveations are necessary before this occurs. Scenes with a limited number of cues relevant to the vision task and which feature a small size×required resolution product will not incur extensive refoveation. If the system had to localize 256 unresolved pixel targets uniformly distributed in a 512 × 512 pixel scene(this distribution places the targets 32 pixels apart in x and y, preventing the 16×16 fovea from registering multiple targets simultaneously), the break-even point would be crossed. However, this

scene does not have few localizations of relevance, which is necessary if foveal vision is to be recommended. In applications where the polygon is built from a low level integrated perception, the size of the polygon is always less after n foveations than the size of n polygons from discrete frames.

As discussed in Section 3.4.10, only those sensor movements within a field-of-view compensating for the lower peripheral acuity should be considered in a comparison between foveal and uniform systems. Other types of spatiotemporal resolution allocation are common to both, such as spotlight search strategies, motion tracking, and fixed gazing (even on a moving target) to compensate for noisy environment (temporal filtering). In fact, these common processes may be more cost-effective in the foveal system than in the uniform system. For example, under extremely noisy conditions, averaging foveal frames involves less bandwidth than averaging uniform frames. Furthermore, the spatial averaging of rexels inherently attenuates uncorrelated noise.

It is difficult to analytically predict the number of entries in a service request map for scenes and tasks of any reasonable complexity. The experiments in this chapter, as well as the unresolved target localization experiments, required only a few foveations. One approach to estimating the number of foveations required to process an object is to analyze the distribution in space and resolution of features in the hierarchical object model, scaled by the expected object scaling in the scene. Only object features relevant to the task are considered. This pruned distribution of the object model is then compared to the foveal polygon of the sensor geometry employed. Effectively, one tries to cover the object model with the minimum number of foveal polygons, each representing a saccade.

Physiological experiments on attention allocation have been conducted which monitor gaze angles and provide some statistical figures for the number of saccades required to accomplish different tasks with static scenes. Human saccades follow a similar strategy to that proposed for foveal machine vision. The fovea is centered on key relevant points in the scene; the greater the localization of task specific relevance in the scene, the fewer the number of discrete positions [Yarbus67]. In some cases, there are only a few key locations in the scene which are interrogated (Yarbus illustrates how foveations on the face of a girl concentrate primarily on the eyes, and to a lesser extent the nose and mouth, as in the Encarnita sequence of Chapter 4). Time permitting, the human observer typically performs a survey mode sequence of foveations where the scene at large is registered.

The human eye is continuously undergoing saccadic motion, at about three to four saccades per second. If the focal plane projection remains constant, temporal filtering will block features after one to three seconds. Also, the time required to integrate a feature into short term memory is typically longer than the average saccade, and the integration process is degraded by premature loss of stimulus [Aaron67]. Consequently, over time, a point of interest in the scene will be revisited many times, but these points can be small in number. This is one opportunity for machine vision to outperform biological vision in static scenes; once a location is registered, the machine should not have to register it again. Of course, the comparison is unfair because the biological approach is structured for dynamic vision. Under time varying scenes, the machine vision system will have to revisit features. Here, foveal vision (machine and biological) is particularly attractive over uniform sampling systems because it offers selective bandwidth reduction given that some minimum frame rate must be maintained.

One limitation of the foveal system with respect to pyramids is in the general inability to dynamically modify the bottom-up algorithm which generates the foveal polygon. Specifically, the bottom-up procedure used to compute a polygon cell must be the same as the procedure used by the foveal sensor to generate a rexel value. Otherwise, the cell values of the foveal manifold (rexels) will differ qualitatively from the computed cells of the polygon, and the individual polygon levels are no longer space invariant representations of the scene.

Foveal sensors can be manufactured with a nonuniform subdivision and an overlapping rexel coverage (e.g., through optical wavefront curvature), thus realizing different values for a and m (6-1). Different convolution functions can be implemented optically, such as with amacronic optics [Lubkin90]. Some of the foveal sensor realization schemes presented in Chapter 7 can also implement negative kernel values and nonlinear functions such as the maximum or median value operator. The foveal sensor can thus support a wide family of bottom-up algorithms. However, once the sensor is "hard wired", so is the algorithm. This excludes the use of feature specific bottom-up generation algorithms unless the vision task is interested exclusively in that one feature, which is not often the case in active vision [Hason88], [Rosen86].

Feature pyramids and integration pyramids are two novel classes of pyramids for active vision [Burt88], [Kjell86], [Zimme86]. The cell values of feature pyramids represent feature presence score functions at different scales and resolutions. Integration

pyramids are generated by nonlinear bottom-up algorithms such as maximum or square value operators which preserve feature cues at higher levels (as opposed to being washed out by linear averaging). A hierarchical set of these pyramids may be employed by the vision system. The generation of these pyramids ultimately begin from Gaussian or Laplacian pyramids whose levels are filtered by a template of the feature of interest. Feature and integration pyramids are supported by foveal sensor data, but only within the boundaries the original Gaussian or Laplacian polygon.

Chapter 7

Conclusions and Directions for Future Work

7.1 Conclusions

The objective of this dissertation was to investigate the properties of active machine vision systems featuring the following attributes found in advanced biological vision:

1. Articulation of the sensor optical axis (a common feature in active vision).
2. Space variant sampling with a non-uniform fixed geometry sensor.
3. Context sensitive data acquisition (dynamic control of sensor optical axis).
3. Processing of signals and integrated perceptions with variable localized bandwidth.

Because of the second attribute, these vision systems are called foveal systems. A principal conclusion is that the performance of the foveal system relative to a uniform acuity system is determined strongly by the vision task and environment. Foveal systems perform best when the bandwidth of features in the scene relevant to the task is localized.

One measure of performance is the amount of processing required to complete a particular task. Reducing this measure signifies the system is extracting more information from less data, and thus supports improvements in the following engineering figures of merit:

1. The vision system can be implemented with smaller hardware.
2. A given processing capability can perform the task in less time.
3. More complex algorithms can be supported while maintaining real-time performance.

4. Greater field-of-view and acuity are supported for a given processing capability while maintaining real-time performance.

Foveal systems have the unique ability to allocate combined spatial and temporal resolution resources to critical regions within the system's field-of-view. This capability is not possible with uniform resolution systems. As a consequence, the computational and data bandwidths of foveal systems have been measured to be several orders of magnitude below that of conventional uniform resolution vision systems performing the same task. These savings can in turn be used to commensurately improve the aforementioned figures of merit. For example, a foveal system can have the same computational and data bandwidths of a uniform system, but offer increased field-of-view and subpixel resolution at the fovea.

Of the foveal geometries studied, the linear and the subdivided exponential offer the greatest potential. The former, which most resembles human acuity, has smoothly graded resolution rolloff and is well suited to classification of extended objects and detection or tracking in a dense multitarget environment. The exponential foveal lattice has sharper discontinuities in resolution, but because these discontinuities are powers of two, the sensor data lends itself to multilevel hierarchical processing better than the other lattices. By subdividing rexels and rexel rings, the acuity profile of the sensor can be customized while retaining the benefits of the exponential geometry. Furthermore, the subdivided geometry is endowed with significant regions of uniform resolution where conventional signal processing can be employed.

The amount of data in a uniform sensor frame is the product of the area of the field-of-view $N \times N$ and the resolution of the system. In this work, the unit of measurement was normalized to the pixel itself (i.e., resolution equals 1). Consequently, the uniform sensor frame contains N^2 values. The frame from a foveal sensor with linear geometry and the same field-of-view and maximum resolution (at the fovea) has $2N$ data values. The significant savings in data over uniform frames grow monotonically with field-of-view and are obtained from the reduced acuity of the foveal sensor at the periphery. A linear foveal sensor with the same amount of data as in the uniform sensor has a resolution of factor $\frac{N}{2}$ greater, or a field-of-view of size $\frac{N^2}{2} \times \frac{N^2}{2}$.

The frame from a foveal sensor with exponential geometry and the same field-of-view and maximum resolution (at the fovea) has $12\log_2 N$ data values. The savings in data

over uniform frames is more significant with the exponential geometry than with the linear geometry because peripheral acuity in the former decreases at a greater rate with distance from the fovea. A linear foveal sensor with the same amount of data as in the uniform sensor has a resolution of factor $\frac{N^2}{2\log_2 N}$ greater, or a field-of-view of size $2^{\frac{N^2}{12}} \times 2^{\frac{N^2}{12}}$.

The significant savings possible in the of number sensor elements in a foveal focal plane array is particularly attractive in applications involving extremely large fields-of-view and small targets. For example, some space defense applications require infrared sensors with over 20 million elements, whereas current manufacturing techniques can approach only 3% of such a pixel density [Bailey88], [Martin87]. Sensor technology and vision processing have been identified as deficient for these applications [AIAA89], [USC87], [USC89].

The read-out device is one limiting factor in such large arrays, imposing a pixel density/frame rate trade-off [BurtD87]. Another limiting factor is sensor cooling. Cryogenic systems can require from 100 to 5000 watts of input power to eliminate one watt of FPA dissipated heat depending on the sensor operating temperature (100 watts/watt for 70° K, 5000 watts/watt for 10° K) [Steln89], [Wolfe85]. At 30mW dissipation for a 256×256 array, this continuous overhead is very costly in space based platforms where power sources are small and expensive.

A complexity measure was computed, consisting of the product of data structure size and total number of time steps required to complete the vision task. The complexity measures for the task of unresolved target localization within an $N \times N$ pixel scene are given in Table 7.1-1 for various conventional systems and different implementations of foveal systems. Due to the frame data savings, foveal systems offer the smallest data structure size. Even for larger fields-of-view, where peripheral acuity is lower, only a few saccades were required to direct the fovea at the target. The performance (saccadic localization error) of the foveal machine vision system using the linear sensor geometry is similar to that of the human visual system performing the same task.

Several techniques for the fusion of data from multiple sensor frames are presented. The resulting integrated perceptions represent the local acuity or ambiguity of information within the global knowledge base. One technique for static scene perception, newly defined in this dissertation, works at the level of measurement data, whereby sensor frames are merged into one larger effective frame which becomes the system's integrated

perception of the scene. This technique is called the discard method, because it drops the lower acuity data when the fields-of-view of two sensor frames overlap. A two dimensional data structure is thus formed with the highest acuity measurement made over time of any point in the field-of-regard. The overall growth rate of the data structure is no greater than $O[\sqrt{n}]$, as opposed to $O[n]$ with multiple frame buffering, where n is the number of frames processed. The information dropped by the discard method is negligible since data is replaced with higher acuity measurements. It is shown that for both synthetic and natural scenes, typical RMS error in the reconstruction of a previously integrated frame can be well below 0.1%, and even less than 0.002% if the irrelevant scene objects not registered with the fovea feature low bandwidth.

Higher level fusion techniques process frames individually and fuse feature or symbolic information. Such an integrated perception is used in several foveal system simulations. Higher level perceptions are not affected by the exact format of the sensor data, but are affected by its information content.

Vision System	Average complexity measure	Worst case complexity measure
Uniresolution uniprocessor	$O[N^4]$	$O[N^4]$
Uniresolution multiprocessor	$O[N^3]$	$O[N^3]$
Gaussian pyramid	$O[N^2\log_2 N]$	$O[N^2\log_2 N]$
Linear foveal resolution uniprocessor	$O[N^2]$	$O[N^2\log_2(\log_2 N)]$
Exponential foveal resolution uniprocessor	$O[(\log_2 N)^3]$	$O[(\log_2 N)^3]$
Linear foveal resolution multiprocessor	$O[N^{1.5}]$	$O[N^{1.5}\log_2(\log_2 N)]$
Exponential foveal resolution multiprocessor	$O[\sqrt{\log_2 N}(\log_2 N)^2]$	$O[\sqrt{\log_2 N}(\log_2 N)^2]$

Table 7.1-1 Average and maximum complexity measures for various machine vision systems performing target localization. The complexity measure is the product of data structure size and the number of time steps required to complete the task.

The foveal vision system generates feature hypotheses based on poorly resolved cues in the integrated perception. Saccadic gaze control is driven by the integrated perception to obtain the information required to confirm or deny the hypotheses. Two saccadic strategies are presented: survey mode and interrogation mode. The first strategy emphasizes global learning, and attempts to minimize the overall entropy of the collective set of hypotheses. The resulting foveations are to unresolved regions where additional cues are expected (hypothesized). The second strategy attempts to maximize the likelihood ratio of a particular hypothesis. The resulting foveations in this case are directly to the cue being interrogated. The survey mode is proposed as the initial gaze control strategy in scene understanding, where a perception of cues (global scene context) is obtained with minimum foveations. The foveal system switches to the interrogation mode when particularly attractive cues have been uncovered and need to be more precisely resolved.

The foveal polygon, newly defined in this dissertation, may serve as a bridge between space-variant sampling lattices and the hierarchical data structures called image pyramids. The polygon is the subset of a pyramid data structure supported by a foveal sensor frame or foveal integrated perception. Coarse-to-fine pyramid algorithms operate within a hierarchy of uniform data levels, thus supporting space invariant algorithms at each level. The "homing-in" property of these algorithms solve the gaze control problem concurrently with image analysis, as in the vertebrate vision system [Weems86]. Queueing algorithms can prioritize multiple visual cues detected in complex scenes for gaze control. Also, multiple cues can be serviced simultaneously with a single foveation. Simulations illustrate that, as with unresolved target localization, very few foveations are required to process scenes with reasonably scaled objects of relevance to the vision task. To preserve shift invariance at polygon levels, the fine-to-coarse algorithms directly manipulating the sensor data must match the rexel generation scheme of the sensor. Foveal sensors lend themselves to hierarchical lowpass filtering (Gaussian polygon generation). Laplacian, feature, and integration pyramids can then be generated from these levels.

Just as the reduced peripheral acuity dramatically reduces the data size in foveal frames with respect to uniform frames with the same field-of-view and maximum resolution, it reduces the size of the polygon data structure with respect to similarly normalized pyramids. The difference between polygon and pyramid data structure also monotonically increases with field-of-view. In fields-of-view between 128×128 and 2048×2048 pixels, the ratio of pyramid to polygon data structure size increases from two to

four orders of magnitude. When a hierarchical multiprocessor architecture is employed, these savings in data structure translate to savings in processing hardware.

7.2 Topics for Further Research

The deliberation on all machine vision topics in the context of foveal systems is of course outside the scope of this dissertation. Important topics, including active vision in three dimensions, dynamic vision, and foveal sensor implementations need further research. The following sections give some preliminary thoughts on these topics.

7.2.1 Active Vision with Foveal Systems

Foveal systems inherently fall into the realm of active vision by the very nature of resolution allocation and gaze control. This effort has concentrated on foveal and active vision in two dimensions, where the sensor axis is translated in a planar fashion in front of a focal plane, or gimbaled in front of a distant focal plane. Foveal systems readily support operation in three dimensional space, as was illustrated by the scale invariance maintained in the resolved object image processing exercises of Chapter 6. However, there seem to be little results in the topic of hierarchical models (in either 2-D or 3-D) [Uhr87]. Such work would be directly applicable to foveal vision and polygon processing, and to hierarchical vision in general.

Further research is required on how the graded resolution of foveal sensors can be exploited by depth perception and both monocular and binocular focusing. Foveations would now be preformed in three dimensions, with foveations along the focal plane incurring sensor translation or gimbaling, and foveations along the depth axis incurring sensor refocusing. In some cases, the foveation strategy may have to optimally service cues requiring further resolution in the three dimensions (e.g., interrogating a long feature radially positioned with respect to the camera). The concept of reduced bandwidth frame data should also be applied to hierarchical approaches other than the rudimentary Gaussian and Laplacian pyramids.

The refinement of sensor positioning beyond that performed by uniform acuity systems may give the foveal system an edge in the implementation of translation sensitive algorithms. For example, foveal systems may be efficient platforms for optical transforms and algorithms based on the Mellin transform due to centering of the feature of interest in the field-of-view by gaze control [Casas76], [Casas77], [McMil87]. This work primarily addressed saccadic sensor motion, but in general gaze control strategies should integrate body motion and robotic control [Lehma83], [Perci88]. These issues remain to be addressed in the context of foveal systems and foveal gaze control. It would also be interesting to see how models for biological performance features could be implemented in foveal machine vision systems, including hyperacuity and superresolution [Burbec87], [Steve89]. It has already been seen in this work that the linear geometry foveal machine vision system has both static and dynamic similarities with human vision. A non-uniformly subdivided exponential pattern approximating the linear acuity profile can be used to support additional experiments using polygons and resolved objects.

It is interesting to note that polar versions of the linear and exponential acuity roll-off geometries have been proposed for polar addressed pyramids [Peleg87]. The geometries were not proposed as sensor topologies but as techniques for resampling uniform resolution frames, so the savings in bandwidth were not extended through the vision system to the sensor. Nevertheless, the concepts of adaptive non-uniform resampling was discussed, providing the necessary "hooks" in the pyramid processing for a gaze control strategy. The recently constructed linear polar FPA would lend itself to this design, except possibly for the uniform Cartesian fovea (essentially a small uniform FPA within a larger polar FPA) [Vander89]. The resulting pyramid, however, does not use Cartesian space invariant image processing techniques, and suffers from interpolation error at every process of level generation.

7.2.2 Dynamic Vision with Foveal Systems

Dynamic imagery refers to sequences of foveal images in which the scene intensities are changing over time intervals of the same order or less than the frame rate. This can arise as a consequence of objects moving in the scene, rapidly evolving clutter, complex movement (other than translation along the focal plane) of the sensor platform

relative to the scene, or in an otherwise static situation simply by the addition of large amounts of sensor noise which cause the scene to vary significantly from perception to perception. In these environments, integrated perceptions cannot be generated from the straightforward fusing of rexels from different frames (e.g., the discard method) because the frames do not necessarily correlate even where they overlap. Higher level perceptions are generated after accounting for motion.

Methods of machine processing of dynamic imagery fall into two categories: those derived from intensity-based optical flow calculations, and feature-displacement derived correspondence methods. In the former case estimates of the spatio-temporal derivatives of image intensity across the entire image are used to infer motion across the image plane and ultimately in three spatial dimensions, while in the latter case key points on key objects within the image are detected, classified and labeled and the displacements of these points used to quantify motion. Optical flow seems indicated when there is very little world knowledge concerning key objects populating the scene, and when simple decisions must be made very rapidly. Indeed there is strong evidence for the existence of optical flow type processing in low level biological vision, for instance to trigger the "looming reflex" in which an object apparently closing rapidly is reflexively avoided without any effort to identify the object or its features as would be required for correspondence processing.

An approach to motion measurement somewhat related to the optical flow category is straightforward image subtraction (temporal differencing). This approach is often used for static clutter/background rejection.

For other more model based scenarios, the preprocessing required for correspondence methods pays off in direct motion estimation on selected objects of interest. In the human visual system, it appears that high level motion tasks such as tracking of selected objects in busy dynamic environments is achieved using this type of knowledge based processing [Ullma87]. An important topic for further research is the analysis of dynamic foveal polygons employing dynamic pyramid algorithms of both categories [Anand87], [Dengler86], [Vergh89].

Foveal vision has the unique property over uniform acuity systems of naturally following perspective distortion and the focus of expansion, assuming that the acuity profile is unimodal with highest acuity at the optical axis. The focus of expansion is the point generated by perspective distortion at infinity from which the scene seems to be expanding as the observer moves (Figure 7.2.2-1). In applications where the vision

platform is moving in an environment cluttered with obstacles, higher acuity directed at the focus of expansion may be desired to resolve objects and possible collisions much sooner than if acuity were spread uniformly throughout the field-of-view. The lower foveal acuity at the periphery of the field-of-view is appropriate as objects that are closer and being successfully avoided appear larger and at the periphery. Thus, foveal systems do not suffer from undersampling at the focus of expansion nor oversampling at near field to the degree of moving uniform acuity systems.

Stationary objects with which the system is on a collision course, or where both the vision platform and the object are moving along collinear trajectories (e.g., a chase or a fight), will be greatly resolved at the fovea. Consequently, foveal systems can be particularly appropriate for homing-in applications. It is interesting to observe how the unimodal acuity roll-off works in conjunction with perspective distortion in an attempt to provide somewhat uniform acuity in all three spatial dimensions, where the depth axis is along the path of motion. One wonders if this was not a key "requirement specification" for biological vision.

This appropriate matching of acuity to perspective distortion supports flow field calculations by better resolving slow moving cues at the focus of expansion which might otherwise not be considered. Of course, foveal vision supports feature correspondence by allocating resolution resources to those features of relevance to the task.

Tracking is another dynamic vision task for which foveal vision is well suited. Here, the sensor axis should be centered at the object being tracked (or the tracking gate center, which may possibly be offset by some lead pursuit strategy). The higher acuity at the fovea permits accurate velocity estimation and good lock-on properties. Should the object move from the axis, due to tracking errors or intentional refoveation to other cues, the lower acuity perception provides centroid information for reaquisition of lock. The closer the object is to the fovea, the more precise is the object position estimate, as was observed in the card exercise on Chapter 6. Therefore, more foveations may likely be required to reacquire lock if the object has been perceived for a long period of time (wide state propagation variance) by large rexels (wide measurement variance). However, the key point is that the object remains in the wide field-of-view of the sensor, which is made possible by the lower peripheral acuity. This is very significant because the probability of loss of acquisition jumps significantly when the object is no longer being perceived, since there is no measurement data to impose an upper bound on location estimate variance, and the cost to redetect and reaquire an object is very costly [Black87].

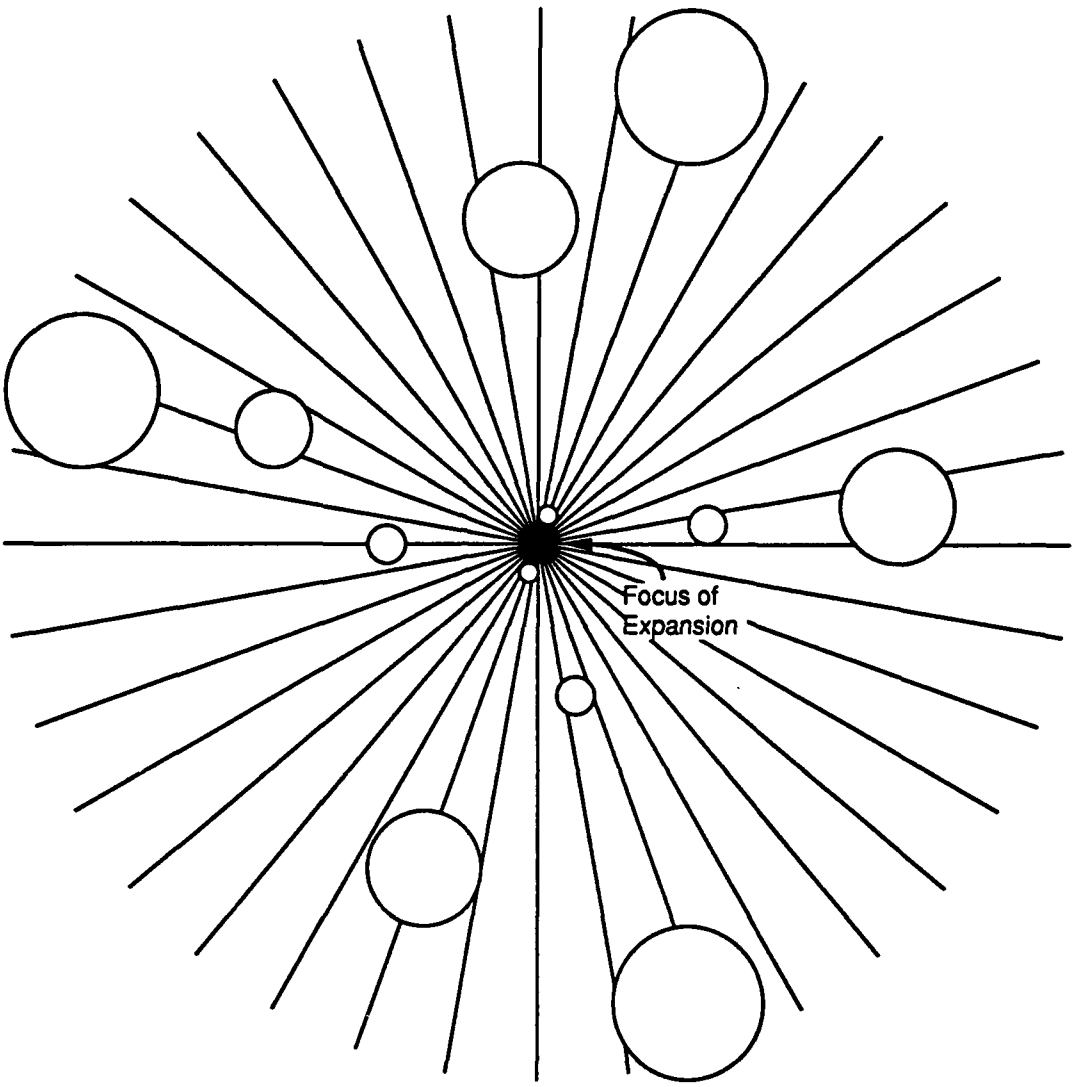


Figure 7.2.3.2-1. Perspective distortion and focus of expansion. All spheres are of the same physical size, but the view is scaled by perspective distortion.

The applicability and resulting performance of existing target detection and tracking algorithms to foveal vision systems should be investigated [Casas86], [Grunin88], [Mao88], [Shao88], [Weiss88].

7.2.3 Foveal Sensor Implementations

A variety of implementation strategies exist which appear capable of producing foveal sensors. Methods for achieving graded resolution are considered separately from those for achieving foveal axis pointing. While most graded resolution strategies are consistent with most pointing strategies, these two basic capabilities must be coupled at some engineering stage. For instance, a bulky, heavy sensor head would not fit well with a pointing system based on moving the entire sensor in order to align the focal plane array to its required foveal axis.

7.2.3.1 Implementation by Software

A solid state camera containing a standard uniform CCD or CID chip, conventional lens, electronics and frame store could be used to acquire conventional pixel data. The appropriate pixels defining each rexel in the foveal array may be summed, scaled and quantized, yielding the desired foveal data array stored in the computer memory. The combining operations may be implemented by a separate coprocessor or array processor communicating with the frame store through a fast image data bus, or by the main system CPU. In either event the foveal lattice is mapped in software (or firmware) rather than hardware.

The obvious advantages of a software implementation are first that it uses completely off-the-shelf technology which would absorb minimal cost and development time, and second that a software foveal lattice can be instantly modified, either to compare selected lattice geometries or to be modified data-adaptively in real-time yielding a dynamically reconfigurable foveal lattice not feasible under any other implementation strategy. This approach permits the use of any bottom-up pyramid algorithm for the computation of a rexel value, since the full array of pixel data (i.e., k=0) is available. Different algorithms may be switched in conjunction with the polygon for different frames.

The limitations of this strategy are perhaps even more apparent. First, all components of the vision system ahead of the pixel to rexel averaging fail to exploit the

advantages of a greatly reduced number of sensor elements and frame data. Pixel frame buffers and busses would be orders of magnitude greater than if the sensor directly provided rexel data. Second, the higher front end bandwidth requirements constrain the frame rate and size, limiting the performance of the overall system.

This implementation is of value in that it can serve as a test bed for the study of foveal system characteristics in a simple and inexpensive laboratory environment. This is a useful first step in the development life cycle from pure digital simulation to full hardware prototyping. Design optimization and performance studies can be carried out by altering the foveal geometry and the follow on data processing steps conveniently, while the system continues to operate in the context of real image data. It was in this fashion that the simulations in this effort were performed: rexel frames were generated by first uniformly digitizing the scene at fovea resolution and then performing pixel averaging.

7.2.3.2 Implementation by Optics

The laboratory convenience of being able to use conventional solid state cameras with their uniform sensor may also be secured by optical design. Consider the lens illustrated in Figure 7.2.3.2-1. This lens is ground to have the (peculiar) property of creating a real focused image not across a plane orthogonal to its optical axis, but on a curved focal surface behind the sensor. Such aspheric lens design with specified wavefront distortion is difficult but not impossible. The lens may be ground from a single blank, built up in layers of thin annular rings cemented together to produce the desired focal surface, etched as a distorted Fresnel lens, or implemented with amacronic optics [Lubkin90].

The rays propagating toward the focal surface are intercepted by the photodetector array. With proper choice of the focal surface geometry, the lattice sites of the uniform photodetector array may be made to transduce exactly the rays which would impinge on the corresponding rexel in an associated foveal lattice geometry. For instance, a parabolic focal surface geometry is equivalent to the linear fovea lattice. Thus the rexel intensities have been distorted to a pixel lattice optically and measured by a conventional sensor array.

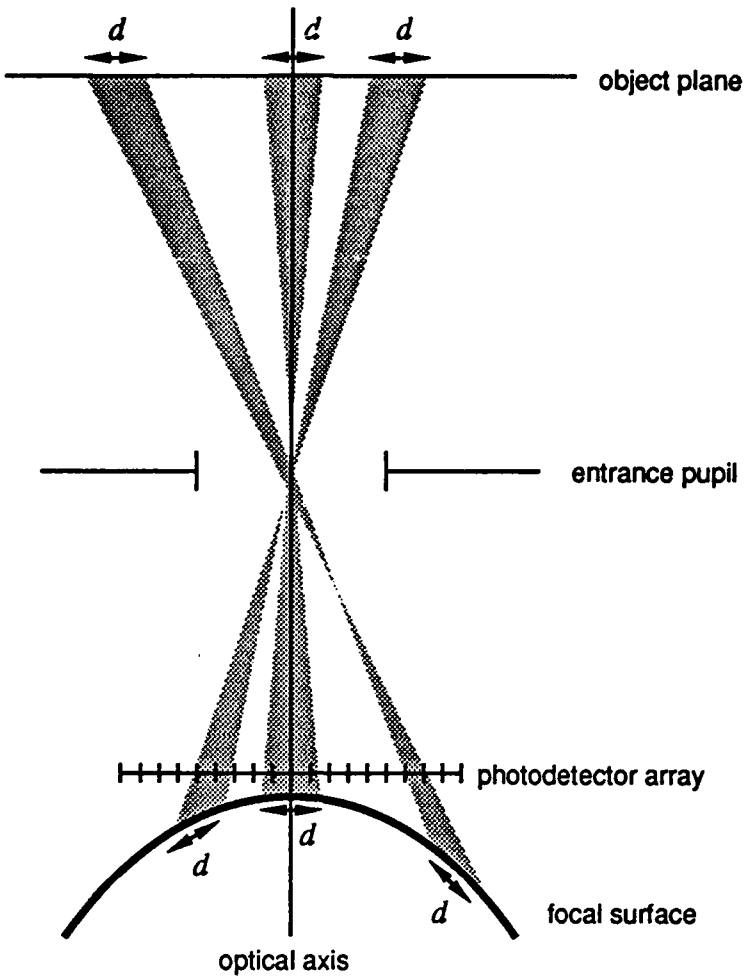


Figure 7.2.3.2-1. Optical implementation of a foveal sensor.

Advantages of this implementation include use of standard solid state cameras and front-end electronics, and the fact that the sensor has only as many pixels as the foveal pattern has rexels. The pixel-to-rexel data compression is effectively accomplished in the analog domain and need not be supported by electronics. The optical implementation, unlike the software implementation, exhibits the principal geometric feature of machine foveal vision: single detectors integrating rays from larger solid angles as eccentricity (distance from foveal axis) increases.

Principal disadvantages include the lens development time and cost, the ultimate accuracy of the wavefront shaping, and the inherent misuse of dynamic range. Pixels on the periphery will capture rays from a much larger solid angle than pixels in the central

zone, and thus a strong intensity gradient will exist across the sensor array when it is observing a relatively uniformly illuminated scene. This may perhaps be compensated by a carefully graded coating of the lens, reducing the transmission to the periphery of the sensor array to that of the center.

As an alternative to conventional lenses, holographic lenses should be considered. While not as efficient as glass lenses, they can be computer-drawn to closely approximate arbitrary lossy transmission functions including phase shift and attenuation. For custom aspheric designs of the type required here, holographic lenses should be considered.

7.2.3.3 Implementation by Monolithic Combiner Circuitry

This approach is a dedicated hardware version of the pixel processing software implementation described above. Here, the sensor chip is implemented with a multilevel VLSI chip supporting conventional uniform lattice of photodetectors and a circuit layer of analog combine circuits between the sensor sites and the readout logic (Figure 7.2.3.3-1). The combine circuits average the pixel samples into rexel values which are then conveyed by the readout logic. The topology of these circuits is determined by the foveal geometry. Any pyramid bottom-up generation algorithm may be hard wired.

The monolithic approach has neither the off-the-shelf-technology advantage nor the reprogrammability of the software approach, but shares with the optical implementation the important feature that rexel integration is done in the analog domain, thus in parallel with discrete time signal conditioning and digital signal processing, and without using clock cycles. However, because rexels are generated behind the photodetectors as with the software implementation, many more individual photodetectors than rexels are still required. A feature of the monolithic approach is that the VLSI chip design may be undertaken using the existing manufacturing techniques.

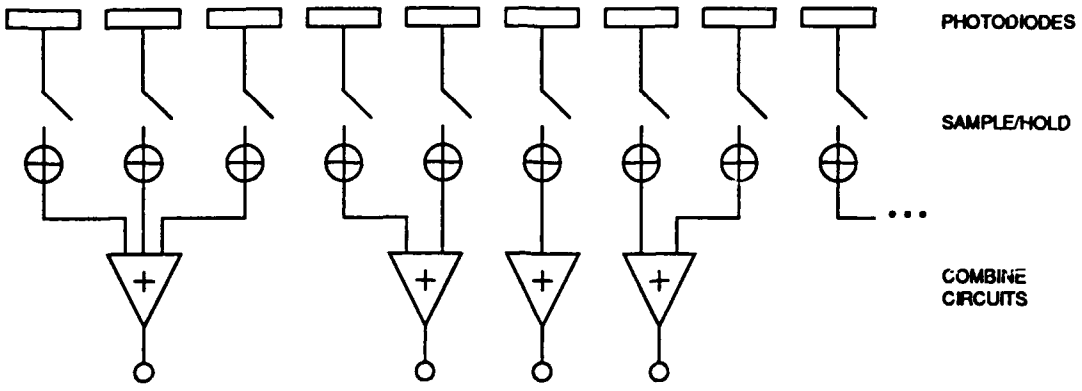


Figure 7.2.3.3-1. Monolithic combiner circuit implementation of a foveal sensor.

7.2.3.4 Direct VLSI Implementation of Foveal Geometry

Perhaps the most direct transcription of the foveal pattern into hardware is obtained by laying out a mask with foveal geometry for the surface layer of a VLSI chip and etching a grid of photodetectors accordingly. Standard line-transfer readout technology could be modified for this design by laying the readout paths radially as shown in Figure 7.2.3.4-1. Thus each line of data would consist of sequences of rexels of every size, one per ring. Conventional optics and frame stores (albeit with new addressing conventions) could be used. This is the only true rexel-based sensor among the suggested implementations.

The principal advantage of the rexel layout VLSI chip implementation is its combination of structural simplicity and full functionality. Its principal drawback may be the novel aspects of a chip of this kind and the probable development costs and time. Among the approaches presented, this implementation is likely to least exploit off-the-shelf technology.

While this is perhaps the cleanest implementation, in the sense that full foveal functionality is supported with conventional optics and the lowest device count, it is arguably not the closest to the biological prototype. This distinction is probably earned by the previous combine circuit layer implementation, in which the outputs of identical photodetectors (rods and cones in the biological model) are combined in increasing number in the periphery of the array (retina). There are clear practical advantages to making all photodetectors of the same size, even perhaps at the price of additional combine circuits.

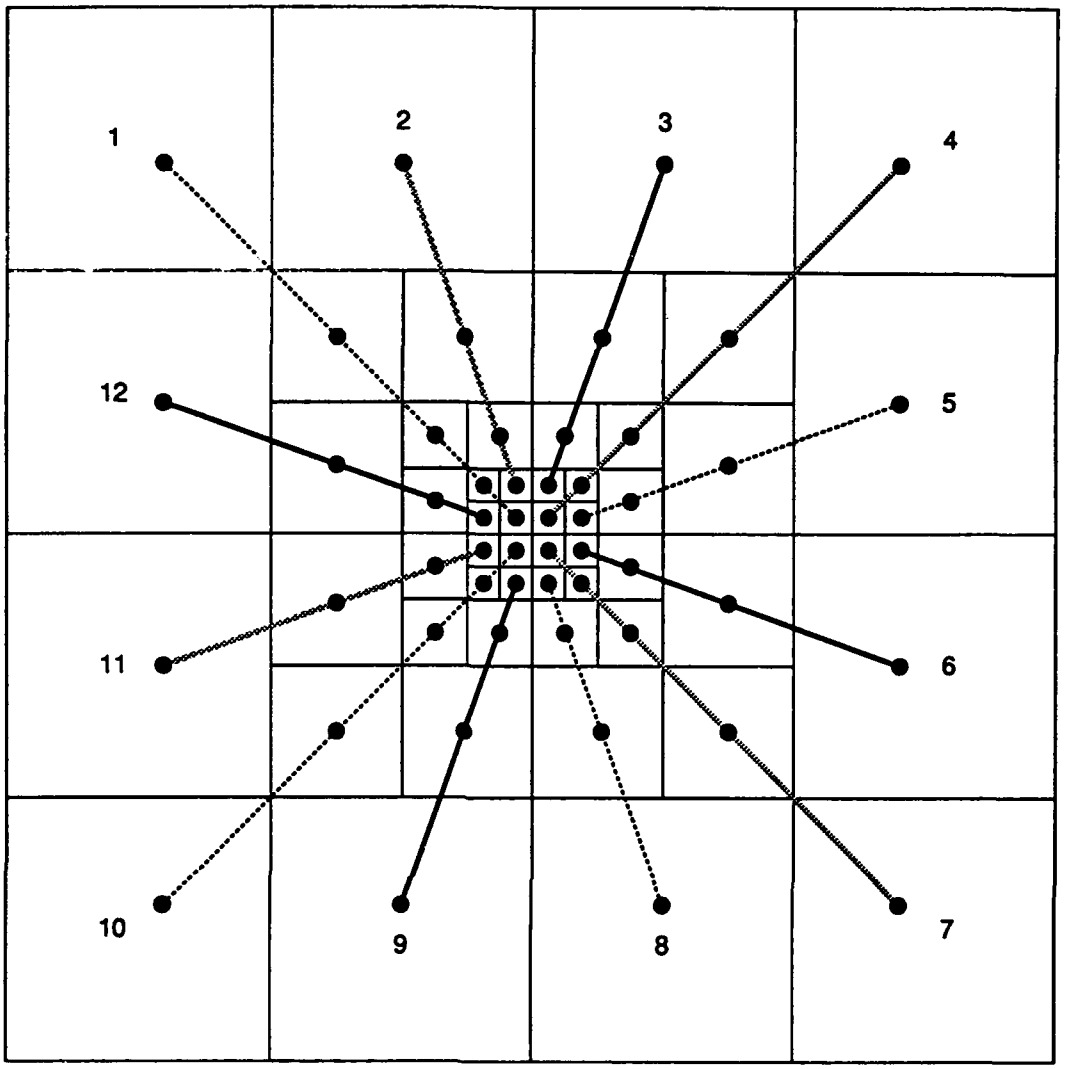


Figure 7.2.3.4-1. VLSI FPA implementation of a foveal sensor.

7.2.3.4 Sensor Pointing Mechanisms

One set of approaches to the implementing of a gazing capability may be characterized by keeping the sensor array stationary and deflecting the incoming light. The most promising technology for accomplishing this is by placing a mirror in the optical path, and precisely realigning the mirror to bring a different portion of the field-of-regard onto the sensor. Great strides have been reported in the design of very fast high precision mirror systems, primarily for missile seeker applications.

A second strategy based on deflecting the light before it reaches the sensor is to use a liquid crystal light valve of sufficient resolution to deflect the light with high angular accuracy. While probably not as accurate as a high precision mirror system, the computer-controlled light valve is lighter, faster and requires no counterbalancing.

Alternatively, the sensor array itself may be moved. To minimize pointing time, power draw and motor ratings, the mass and size of the portion of the sensor module being moved should be minimized. The need for bulky cryogenics with infrared sensor arrays complicates the approach for such applications. Another consideration is that the analog to digital conversion hardware must be situated closely to the sensor array in order to minimize path length of the low level sensor signal and its associated maladies (crosstalk, noise).

The pointing strategy chosen by evolutionary "consensus" is to separate out the lens, sensor array and conversion systems onto a light sensor head (the eyeball) which is moved relative to the scene while the heavier elements of the system remain stationary. Great pointing accuracy is more difficult to achieve in such a system than by moving a lighter mirror, but the relatively large central zone of the human retina (perhaps one degree cone angle) makes great accuracy unnecessary. While the evolutionary strategies are excellent starting points for considering machine vision implementations, the conditions in which organisms must survive, driving factors in biological system evolution, are typically quite different from the conditions in which a machine vision system must function.

7.2.3 Foveal Processing of Object Oriented Databases

Foveal systems selectively measure a state of nature. In an abstract sense, the same definition applies to database languages. Whereas a large database may contain gigabytes of information, only a few fields on a particular feature may be of relevance to some application. Current efforts in object oriented database construction, particularly in the realm of simulation, are relating feature data to objects just as they are clustered in nature. There is obvious benefits to a database language which in a single interrogation can retrieve most or all of the information about an object, plus less extensive information on surrounding objects so as to provide contextual information relating the principle object

under interrogation to the state of nature as a whole. Throughout this document, acuity has been interpreted in the conventional spatial sense. It is unclear at this point how spatial metrics, including the foveal geometries themselves, map to this environment. Nevertheless, if the bandwidth benefits of foveal vision can be retained, foveal database languages may be of significant value for the processing of large object oriented databases.

7.2.4 Transform Domain Representation of Foveal Images

A transform domain representation of foveal frames and the integrated perception is of value to foveal vision just as transforms are indispensable for uniform resolution image processing. However, conventional transforms, such as the Fourier transform, are defined for space invariant systems, i.e., systems featuring uniform sampling and global bandwidth measures. The transform of foveal data must support non-uniform sampling, and the concept of localized bandwidth [Saleh85].

There are a number of unique aspects of foveal vision which distinguish it from most of the published work on non-uniform sampling. First, the value of a rexel with linear dimensions m, r_m is not an approximation to the convolution of the analog scene $f(x,y)$[24] with a dirac function as a conventional sample s_δ

$$s_\delta(i,j) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(u,v)\delta(u-iT, v-jT)\,du\,dv \tag{7-1}$$

but is instead a space variantly filtered and sampled version of the scene

$$r_m = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(u,v)w_m(u,v)\,du\,dv \tag{7-2}$$

The sampling window represents the rexel value generating kernel, which in our work models the integration of all energy registered by the rexel,

$$w_m(x,y) = \begin{cases} 1 & x \in R_{m,x}, y \in R_{m,y} \\ 0 & \text{otherwise} \end{cases} \tag{7-3}$$

[24] The function $f(x,y)$ is the projection of the actual scene onto the focal plane of the sensor.

where m is the rexel index and $R_{m,x}$, $R_{m,x}$, are the ranges of the rexel coverage in x and y respectively. Other windows besides the rectangular window can be used; depending on the hardware implementation of the sensor, overlapping windows and negative weighting (differencing windows) can be implemented.

A second unique aspect of foveal vision is the desired transform representation. Most of the published work on non-uniform sampling attempts to reconstruct a uniform representation of the signal with a sampling lattice satisfying the Nyquist rate. The objective of a foveal data transform is to retain in the transform space the information density advantages of variable acuity sampling. Also, most of the published work on transforms supporting a localized bandwidth measure employ uniform sampling.

A third unique aspect of foveal vision is the sensor geometry itself. The majority of published work on non-uniform sampling employs either a smooth (i.e., "rubber sheet") distortion of the Cartesian geometry or polar sampling. The foveal geometries analyzed in Chapter 3, and all derivatives from rexel subdividing, feature stepped acuity and are not circularly symmetric.

A transform domain which supports localized bandwidth is the combined spatial frequency domain (referred to from this point on as the CSFD), in which the transform of a signal is a function of both spectral and spatial arguments, as opposed to one or the other (e.g., time domain or Fourier representations of signals). Such a transform requires a basis function which does not have uniform magnitude throughout the spatial domain (hence ruling out $e^{-j\omega x}$), otherwise there would be no spatial selectivity.

One candidate basis function is the Gabor elementary function. This function is a Gaussian enveloped complex sinusoidal, and is expressed in one dimension (the two dimensional function is separable) by

$$\gamma_{m,n}(x) = \left(\frac{\sqrt{2}}{D} e^{-\pi\left(\frac{x-mD}{D}\right)^2} \right) e^{jnWx} \tag{7-4}$$

$$WD \leq 2\pi \tag{7-5}$$

where m and n are integers indices which situate the basis function in the CSFD, and W and D determine the aspect ratio of the energy contour (e.g., in the extreme cases of $D \rightarrow 0$ or $W \rightarrow 0$, the basis function is a dirac function in the spectral or spatial domain,

respectively). When $WD=2\pi$, the Gabor function has the distinct feature of being maximally localized in the CSFD.

A Gabor elementary function represents a spatially tuned filter. The above expression is for uniform sampling systems. Gabor functions have been used to represent non-uniformly sampled images, where the sampling lattice is a uniform grid stretched by a smooth function [Porat88]; the representation stretched the Gabor functions by the same smooth function so as to preserve the general form of the analysis-synthesis pair of the uniform case.

A drawback of the Gabor elementary function is that it is space and band unlimited. Consequently, an infinite set of basis functions are required to represent a bandlimited or spacelimited signal. Furthermore, the "rubber sheet" sampling grid distortion raises the machine vision implementation issues discussed in Section 3.2. Nevertheless, Gabor functions may offer insight the types of vision operations that can be conducted in the CSFD on foveal data. It has been experimentally confirmed that an early "hardwired" step in vertebrate vision is the decomposition of visual signals into spatially tuned channels by filters with kernels resembling Gabor functions [Daugm83], [Hess89], [Krona85], [Malla89], [Polle83], [Shlom87], [Wilson88]. Furthermore, the filters are highly quantized, with information represented primarily by signal phase at a resolution between 15° and 45° (phase being quantized to no more than 24 values) [Bchar88], [Caelli83].

Another candidate basis function is the prolate spheroidal wave function (PSWF), which offers maximum localization in the CSFD with the constraint of being bandlimited [Landa60], [Landa62], [Slepi61], [Slepi64], [Slepi65], [Slepi78]. For the representation of foveal signals, this property represents the major distinction between prolate spheroidal and Gabor functions; a bandlimited signal can be approximated by the combination of a finite number of PSWFs, and this combination will itself be bandlimited. The DPSWF has been employed as a tool for optimal filtering in the localized bandwidth sense, with maximum passband flatness and transition band slope [Bronez88], [Mathe85], [Papou72], [Tufts70].

Most of the literature on CSFD representations assumes uniform sampling. Transforms have been implemented on hierarchical architectures which decompose uniform resolution images into basis functions which are localized in this domain [Adels87]. A hierarchical approach may simplify the process of computing transforms of foveal data in the CSFD, as it did with space domain foveal image processing.

References

[Aaron67] Doris Aaronson, "Temporal factors in perception and short-term memory", in Ralph Norman Haber (ed.), Contemporary Theory and Research in Visual Perception, Holt, Rinehart, and Winston, 1968.

[Adels87] Edward H. Adelson, Eero Simoncelli, Rajesh Hingorani, "Orthogonal pyramid transforms for image coding", *SPIE Visual Communications and Image Processing*, vol. 845, pp. 50-58, 1987.

[AIAA89] American Institute of Aeronautics and Astronautics, workshop program report, *AIAA Assessment of Strategic Defense Initiative Technologies*, AIAA, March 1989.

[Aliom87] J. Aliomonos, A. Badyopadhyay, "Active vision", *Proc. 1st IEEE Int. Conf. on Computer Vision*, pp. 35-54, 1987.

[Anand87] P. Anandan, Mark F. Cullen, "Robust Parallel Computation of Image Displacement Fields", *SPIE*, vol. 848, *Intelligent Robots and Computer Vision*, pp. 248-254, 1987.

[Bailey88] G. C. Bailey, "A 256 by 256 HgCdTe hybrid focal plane for low-background Earth resources and astronomy applications", *Proc. SPIE Infrared Technology XIV*, vol. 972, pp. 122-126, 1988.

[Bajcsy86] Ruzena Bajcsy, "Tactile information processing", in Pyramidal Systems for Computer Vision, V. Cantoni, S. Levialdi (eds.), Springer-Verlag, pp. 341-356, 1986.

[Bajcsy88] _____, "Active perception", *Proc. IEEE*, vol. 76, no. 8, pp. 996-1005, 1988.

[Balla87] Dana H. Ballard, "Eye movements and visual cognition", in *Proc. Workshop on Spatial Reasoning and Multisensor Fusion*, pp. 188-200, 1987.

[Balla89] _____, "Behavioural constraints on animate vision", *Image and Vision Computing*, vol.7, no.1, 1989.

[Baloh88] R. W. Baloh, R. D. Yee, V. Honrubia, K. Lacobson, "A compariason of the dynamics of horizontal and vertical smooth pursuit in normal human subjects", *Aviation, Space, and Environmental Medicine*, Feb., 1988.

[Bchar88] J. Bchar, M. Porat, Y.Y. Zeevi, "The importance of localized phase in vision and image representation", *SPIE*, vol. 1001, *Visual Communications and Image Processing*, pp. 61-68, 1988.

[Bessl86] Ph. W. Besslich, "Pyramidal transforms in image processing and computer vision", in Pyramidal Systems for Computer Vision, V. Cantoni, S. Levialdi (eds.), Springer-Verlag, pp. 215-246, 1986.

[Berge83] James R. Bergen, Bela Julesz, "Rapid Discrimination of Visual Patterns", *IEEE Trans. SMC*, Oct-83, vol. 13, no. 5, pp. 857-863.

[Black87] Samuel S. Blackman, Multitarget tracking with radar applications, Artech Publishers, 1987.

[Blanf88] Ronald P. Blanford, Steven L. Tanimoto, "Bright-Spot Detection in Pyramids", *Computer Vision, Graphics, and Image Processing*, vol. 43, Academic Press, pp. 133-149, 1988.

[Bothw87] M. Bothwell, G. C. Bailey, V.G. Wright, "Short wavelength 128 by 128 focal plane arrays for remote sensing applications", *Proc. SPIE Focal Plane Arrays: Technology and Applications*, vol. 865, pp. 86-91, 1987.

[Bronez88] Thomas P. Bronez, "Spectral Estimation of Irregularly Sampled Multidimensional Processes by Generalized Prolate Shperoidal Sequences", *IEEE Trans. ASSP*, vol. 36, no.12, pp. 1862-1873, 1988.

[Brown89] Christopher M. Brown (ed.), Northeast Artificial Intelligence Consortium Annual Report 1987: Parallel, Structural, and Optimal Techniques in Vision, Syracuse University, Rome Air Development Center report RADC-TR-88-324, March, 1989.

[Burbec87] Christina A. Burbeck, "Position and spatial frequency in large-scale localization judgments", *Vision Res.*, vol. 27, no. 3, pp. 417-427, 1987.

[Burns87] J. Brian Burns, Leslie J. Kitchen, "Rapid recognition out of a large model base using prediction hierarchies and machine parallelism", *SPIE*, vol. 848, *Intelligent Robots and Computer Vision*, pp. 225-233, 1987.

[BurtD87] D. J. Burt, "Read-out devices for focal plane arrays", *Proc. SPIE Focal Plane Arrays: Technology and Applications*, vol. 865, pp. 2-16, 1987.

[Burt83] Peter J. Burt, Edward H. Adelson, "The Laplacian Pyramid as a Compact Image Code", *IEEE Trans. Com*, vol. 31, no. 4, pp. 532-540, 1983.

[Burt84] Peter J. Burt, "The pyramid as a structure for efficient computation", in Multiresolution Image Processing and Analysis, A. Rosenfeld (ed.), pp. 6-35, Springer-Verlag, 1984.

[Burt88] _____, "Smart sensing with a pyramid vision machine", *Proc. IEEE*, vol. 76, no. 8, pp. 1006-1015, 1988.

[Caelli83] Terry Caelli, Martin Hubner, "Coding Images in the Frequency Domain: Filter Design and Energy Processing Characteristics of the Human Visual System", *IEEE Trans. SMC*, vol. 13, no. 5, pp. 1018-1021, 1983.

[Caelli87] Terry Caelli, Mark Nawrot, "Localization of signals in images", *J. Opt. Soc. Am. A.*, vol. 4, no. 12, pp. 2274-2280, 1987.

[Canton86] Virginio Cantoni, "I. P. hierarchical systems: architectural features", in Pyramidal Systems for Computer Vision, V. Cantoni, S. Levialdi (eds.), Springer-Verlag, pp. 31-40, 1986.

[Canton88] Virginio Cantoni, Stefano Levaldi, "Multiprocessor computing for images", *Proc. IEEE*, vol. 76, no. 8, pp. 959-969, 1988.

[Casas76] David Casasent, Demetri Psaltis, "Scale Invariant Optical Transform", *Optical Engineering*, vol. 15, no. 3, pp. 258-261, 1976.

[Casas77] _____, "New Optical Transforms for Pattern Recognition", *Proc. IEEE*, vol. 65, no. 1, pp. 77-83, 1977.

[Casas86] D. Casasent, B.V.K. Vijaya Kumar, Y.L. Lin, "Subpixel target detection and tracking", *Proc. SPIE Intelligent Robots and Computer Vision*, vol. 726, pp. 206-219, 1986.

[Clark84] J. J. Clark, P. D. Lawrence, "A hierarchical image analysis system based on upon oriented zero crossings of bandpass images", in Multiresolution Image Processing and Analysis, A. Rosenfeld (ed.), Springer-Verlag, pp. 6-35, 1984.

[Curla83] John C. Curlander, Vasilis Z. Marmarelis, "Processing of Visual Informio in the Distal Neurons of the Vertebrate Retina", *IEEE Trans. SMC*, vol. 13, no. 5, pp. 934-943.

[Crowl84] J. L. Crowley, "A multiresolution representation for shape", in Multiresolution Image Processing and Analysis, A. Rosenfeld (ed.), Springer-Verlag, pp. 6-35, 1984.

[Daugm83] John G. Daugman, "Six Formal Properties of Two-Dimensional Anisotropic Visual Filters: Structural Principles and Frequency/Orientation Selectivity", *IEEE Trans. SMC*, vol. 13, no. 5, pp. 882-887, 1983.

[Dengler86] Joachim Dengler, "Local motion estimation with the dynamic pyramid", in Pyramidal Systems for Computer Vision, V. Cantoni, S. Levialdi (eds.), Springer-Verlag, pp. 289-298, 1986.

[Dough87] E. R. Dougherty, C.R. Giardina, Matrix Structured Image Processing, Prentice-Hall, 1987.

[Drumm89] Oliver E. Drummond (ed.), Signal and Data Processing of Small Targets 1989, ("Proceedings of a Conference on Digital Signal Processing, Association, and Tracking of a Point Source, Very Small, and Cluster Targets"), Proc. SPIE vol. 1096, March 1989.

[Duda73] Richard O. Duda, Peter E. Hart, Pattern Classification and Scene Analysis, Wiley-Interscience, 1973.

[Duvoi84] H. A. Duvoisin, R. T. Flaherty, W. R. Lawson, "A Systemaic search model incorporating interfixations, glimpse distribution, and clutter", *Proc. IRIS Specialty Group on Imaging*, pp. 27-39, 1984.

[Dyer87] Charles R. Dyer, "Multiscale image understanding", in L. Uhr (ed.), Parallel Vision Systems, Academic Press, pp. 171-214, 1987.

[Eckmil83] Rolf Eckmiller, "Neural Control of Foveal Pursuit Versus Saccadic Eye Movements in Primates - Single-Unit Data and Models", *IEEE Trans. SMC*, vol. 13, no. 5, pp. 980-989, 1983.

[Fisch87] M. A. Fischetti, "The silver screen blossoms into color", *IEEE Spectrum*, vol. 24, no. 8, pp. 50-55, 1987.

[Kunik83] Kunihiko Fukushima, Sei Miyake, Takayuki Ito, "Neocognitron: a neural network model for a mechanism of visual pattern recognition", *IEEE Trans. SMC*, vol. 13, no. 5, pp. 826-834, 1983.

[Grosk84] W. I. Grosky, R. Jain, "Region matching in pyramids for dynamic scene analysis", in Multiresolution Image Processing and Analysis, A. Rosenfeld (ed.), Springer-Verlag, pp. 6-35, 1984.

[Grunin88] J. Gruninger, J. Conant, C. Kolb, "Boost phase missile identification algorithms for SDI using passive IR", (Secret) *Proc. IRIS Targets, Backgrounds, and Discrimination*, vol. 1, pp. 169-201, 1988.

[Hammi90] R. Hamming, Coding and Information Theory, Prentice Hall, 1980.

[Hason88] B. Hason, Y. Y. Zeevi, "Mutual Information of Images: A New Approach to Pyramidal Image Analysis", *SPIE*, vol. 1001, *Visual Communications and Image Processing*, pp. 555-562, 1988.

[He89] Peiyuan He, Eileen Krowler, "The role of location probability in the programming of saccades: implications for 'center of gravity' tendencies", *Vision Research*, vol. 29, no. 9, pp. 1165-1181, 1989.

[Hess89] R. F. Hess, J. S. Pointer, R. J. Watt, "How are spatial filters used in fovea and parafovea?", *J. Opt. Soc. Am. A.*, vol. 6, no. 2, pp. 329-339, 1989.

[Hoffm86] D. D. Hoffman, W. A. Richards, "Parts of recognition", in Alex P. Pentland (ed), From Pixels to Predicates, Ablex, pp. 268-293, 1986

[Jarske88] Petri Jarske, Tapio Saramaki, and Sanjit K. Mitra ,"On Properties and Design of Nonuniformly Spaced Linear Arrays", *IEEE Trans. ASSP*, vol. 33, no. 3, pp. 372-380, 1988.

[Jayant84] N. S. Jayant, P. Noll, Digital Coding of Waveforms, Prentice-Hall, 1984.

[Katsi87] Constantine Katsinis, Alexander Poularikas, "Pattern Recognition with a Spiral Sampling Technique", *SPIE*, vol. 845, *Visual Communications and Image Processing II*, pp. 66-69, 1987.

[Kjell86] Bradley P. Kjell, Charles R. Dyer, "Segmentation of textured images by pyramid linking", in Pyramidal Systems for Computer Vision, V. Cantoni, S. Levialdi (eds.), Springer-Verlag, pp. 273-288, 1986.

[Kreid90] G. Kreider, J. Van der Spiegel, I. Born, C. Claeys, I. Debusschere, G. Sandini, P. Dario, F. Fantini, "A Retina Like Space Variant CCD Sensor", *SPIE*, vol. 1242,*Charge-Coupled Devices and Solid State Optical Sensors*, 1990.

[Krona85] R. E. Kronauer, Y. Y. Zeevi, "Reorganization and diversification of signals in vision", *IEEE Trans. SMC*, vol. 15, no. 1, pp. 91-101, 1985.

[Landa60] H. J. Landau, H.O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty-II", *The Bell System Technical Journal*, pp. 65-84, Jan 1960.

[Landa62] _____, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty-III: The Dimension of the Space of Essentially Time- and Band-Limited Signals", *The Bell System Technical Jounal*, pp. 12954-1336, July 1962.

[Larson74] Larson, H.J., Introduction to Probability and Statistical Inference, Second (ed.), Wiley Series in Probability and Mathematical Statistics, 1974.

[Lehma83] Steven L. Lehman, Lawrence W. Stark, "Perturbation analysis applied to eye, head, and arm movement models", *IEEE Trans. SMC*, vol. 13, no. 5, pp. 972-979, 1983.

[Lemij89] H. G. Lemij, H. Collewijn, "Differences in accuracy of human saccades between stationary and jumping targets", *Vision Research*, vol. 29, no. 12, pp. 1737-1748, 1989.

[Levin85] Martin D. Levine, Vision in Man and Machine, McGraw Hill, 1985.

[Lubkin90] Yale Jay Lubkin, "Smart eyeballs", *Aerospace and Defense Science*, vol. 11, no.5, pp. 11-13, 1990.

[Lynch85] Thomas J. Lynch, Data Compression: Techniques and Applications, Lifetime Learning Publications, 1985.

[Malla89] Stephane G. Mallat, "Multifrequency Chnnel Decompositions of Images and Wavelet Models", *IEEE Trans. ASSP*, vol. 37, no. 12, pp. 2091-2110, 1989

[Mao88] Z. Mao, R. N. Strickland, "Image sequence processing for target estimation in forward looking infrared imagery", *Optical Engineering*, vol. 27, no. 7, pp. 541-549, 1988.

[Mares88] M. Maresca, M. A. Lavin, H. Li, "Parallel architectures for vision", *Proc. IEEE*, vol. 76, no. 8, pp. 970-981, 1988.

[Marr76] D. Marr, "Early processing of visual information", *Philosophical Transactions of the Royal Society, London*, ser. B, vol. 275, pp. 483-524, 1976.

[Martin87] J. Martin, "Sensors for SDI", *Defense Science and Electronics*, vol. 6, no. 5, pp. 36-38, 1987.

[Mathe85] John D. Mathews, J.K. Breakall, and George K. Karawas, "The Discrete Prolate Shperoidal Filter as a Digital Signal Processing Tool", *IEEE Trans. ASSP*, vol. 33, no. 6, pp. 1471-1478, 1985.

[McMil87] John D. McMillen, O. R. Mitchell, "The Orthornormal Rourier-Mellin Transform for Precision Scale Detection and Range Data Acquisition", *SPIE*, vol. 848, *Intelligent Robots and Computer Vision*, pp. 25-32, 1987.

[Mead90] Carver Mead, L. Conway, Introduction to VLSI, Addision-Wesley, 1980.

[Miller88₁] Russ Miller, Quentin F. Stout, "Efficient parallel convex hull algorithms", *IEEE Transactions on Computers*, vol. 37, no. 12, pp. 1605-1618.

[Miller88₂] _____, "Simulating essential pyramids", *IEEE Transactions on Computers*, vol. 37, no. 12, pp. 1642-1648.

[Oguzt83] M. Namik Oguztoreli, "Modeling and simulation of vertebrate primary visual system: basic network", *IEEE Trans. SMC*, vol. 13, no. 5, pp. 766-781, 1983.

[Papou84] Athanasios Papoulis, Probability, Random Variables, and Stochastic Processes, Second Edition, McGraw-Hill, 1984.

[Papou72] Athanasios Papoulis, Miguel S. Bertran, "Digital Filtering and Prolate Functions", *IEEE Transactions on Circuit Theory*, vol. CT-19, no. 6, pp. 674-681, 1972.

[Peleg87] Shmuel Peleg, Orna Federbusch, Robert Hummel, "Custom-made pyramids", in Leonard Uhr (ed.), Parallel Computer Vision, Academic Press, pp. 125-146, 1987.

[Perci88] Lynn C. Percival, Fred E. Guedry, "Performance-based assessment of oculomotor efficiency", *Aviation, Space, and Environmental Medicine*, Feb., 1988.

[Photo89] Photometrics, Charge-Coupled Devices for Quantitative Electronic Imaging, 1989.

[Polle83] Daniel A. Pollen, Steven F. Ronner, "Visual Cortical Neurons as Localized Spatial Frequency Filters", *IEEE Trans. SMC*, vol. 13, no. 5, pp. 907-916, 1983.

[Porat88] Moshe Porat, Yehoshua Y. Zeevi, "The Generalized Gabor Scheme of Image Representation in Biological and Machine Vision", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 1988, vol. 10, no. 4, pp. 452-467.

[Rosen86] Azriel Rosenfeld, "Some pyramid techniques for image segmentation", in Pyramidal Systems for Computer Vision, V. Cantoni, S. Levialdi (eds.), Springer-Verlag, pp. 261-272, 1986.

[Rosen88] _____, "Computer Vision: Basic Principles", *Proc. IEEE*, vol. 76, no. 8, pp. 863-868, 1988.

[Sage77] A. P. Sage, C. C. White, Optimal Systems Control, second edition, Prentice Hall, 1977.

[Saleh85] Bahaa E.A. Saleh, Nikola S. Subotic, "Time-Variant Filtering of Signals in the Mixed Time-Frequency Domain", *IEEE Trans. ASSP*, vol. 33, no. 6, pp. 1479-1485, 1985.

[Schae87] D. H. Schaefer, P. Ho, J. Boyd, C. Vallejos, "The GAM pyramid", in Leonard Uhr (ed.), Parallel Computer Vision, Academic Press, pp. 15-42, 1987.

[Schri88] A. Schrift, Y.Y. Zeevi, M. Porat, "Pyramidal Edge Detection and Image Representation", *SPIE*, vol. 1001, *Visual Communications and Image Processing*, pp. 529-535, 1988.

[Scott90] Peter D. Scott, "Machine vision systems", in N. Liebovic (ed.), Visual Science, Springer Verlag, 1990.

[Shao88] H. M. Shao, T.L. Bergen, "Midcourse tracking by velocity filters", (Secret) *IRIS Proc. Targets, Backgrounds, and Discrimination*, vol. 1, pp. 345-357, 1988.

[Shein84] M. Sheiner, "Multiresolution feature encoding", in Multiresolution Image Processing and Analysis, A. Rosenfeld (ed.), Springer-Verlag, pp. 6-35, 1984.

[Shlom87] E. Shlomot, Y.Y. Zeevi, and W.A. Pearlman, "The Importance of Spatial Frequency and Orientation in Image Decomposition and Coding", *SPIE*, vol. 845 *Visual Communications and Image Processing II*, pp. 152-158, 1987.

[Slepi61] David Slepian, H.O. Pollak, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty-I", *The Bell System Technical Journal*, pp. 43-63, January 1961.

[Slepi64] David Slepian, "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty-IV: Extensions to Many Dimensions; Generalized Prolate Spheroidal Functions", *The Bell System Technical Journal*, pp. 3009-3057, November 1964.

[Slepi65] David Slepian, Estelle Sonnenblick, "Eigenvalues Associated with Prolate Spheroidal Wave Functions of Zero Order", *The Bell System Technical Journal*, pp. 1745-1759, October 1965.

[Slepi78] _____, "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty-V: The Discrete Case", *The Bell System Technical Journal*, vol. 57, no. 5, pp. 1371-1430, 1978.

[Steln89] W. W. Stelner, "Adaptive machine vision annual report for SDI/IST", (Unclassified), SAIC Report under ONR contract N00014-86-C-00601, DTIC #AD-A208 130, 117 pgs., Jan. 1989.

[Steve89] Scott B. Stevenson, Lawrence K. Cormack, Clifton M. Schor, "Hyperacuity, superresolution, and gap resolution in human stereopsis", *Vision Research*, vol. 29, no. 11, 1989.

[Stout86] Quentin F. Stout, "An algorithmic comparison of meshes and pyramids", in Evaluation of Multicomputers for Image Processing, L. Uhr et.al. (eds.), Academic Press, pp. 107-122, 1986.

[Stout87] _____, "Pyramid algorithms optimal for the worst case", in Parallel Vision Systems, L. Uhr (ed.), Academic Press, pp. 147-167, 1987.

[Stout88] _____, "Mapping vision algorithms to parallel architectures", *Proc. IEEE*, vol. 76, no. 8, pp. 982-995, 1988.

[Tanim84] Steven L. Tanimoto, "Sorting, histogramming, and other statistical operations on a pyramid machine", in Multiresolution Image Processing and Analysis, A. Rosenfeld (ed.), Springer-Verlag, pp. 6-35, 1984.

[Tanim87] Steven L. Tanimoto, Terry J. Ligocki, Robert Ling, "A prototype pyramid machine for hierarchical cellular logic", in Leonard Uhr (ed.), Parallel Computer Vision, Academic Press, pp. 43-84, 1987.

[Tufts70] Donald W. Tufts, James T. Francis, "Designing Digital Low-Pass Filterss--Comparison of Some Methods and Criteria", *IEEE Transactions on Audio and Electroacoustics*, vol. AU-18, no. 4, pp. 487-494, 1970.

[Uhr86] L. Uhr, "Parallel, hierarchical software/hardware pyramid architechtures", in Pyramidal Systems for Computer Vision, V. Cantoni, S. Levialdi (eds.), Springer-Verlag, pp. 1-20, 1986.

[Uhr87] _____, "Highly parallel, hierarchical, recognition cones for perceptual structures", in Parallel Vision Systems, L. Uhr (ed.), Academic Press, pp. 249-292, 1987.

[Ullma87] S. Ullman, "Analysis of visual motion by bioligical and computer systems", in Readings in Computer Vision, M. Fischler and O. Firschein (eds.), Kaufmann Publishers, 1987.

[USC87] United States Congress, Office of Technology Assessment, *SDI: Technology, Survivability, and Software*, 100th Congress, Government Printing Office, 1987.

[USC89] _____, Strategic Defense Initiative Organization, *1989 Report to the Congress on the Strategic Defense Initiative*, 100th Congress, Department of Defense, 1989.

[Vander89] J. Van der Spiegel, G. Kreider, C. Claeys, I. Debusschere, G. Sandini, P. Dario, F. Fantini, P. Bellutti, G. Soncini, "A Foveated Retina-Like Sensor using CCD Technology', in Analog VLSI Implementation of Neural Systems, C. Mead and M. Ismail (eds.), Kluwer Acad. Publ., pp. 189-210, 1989.

[Vergh89] Gilbert Verghese, Karey Lynch Gale, Charles R. Dyer, "*Real-time, parrallel motion tracking of three dimensional objects from spatiotemporal image sequences*", in Parallel Algorithms for Machine Intelligence and Pattern Recognition, L. Kanal, V. Kumar, P. Gopalakrishnan (eds.), Springer Verlag, 1989.

[Weems86] Charles C. Weems Jr., "Closing the feedback loop in machine vision", in Evaluation of Multicomputers for Image Processing, L. Uhr et.al. (eds.), Academic Press, pp. 161-180, 1986.

[Weiss88] A. J. Weiss, B. Friedlander, W.G. Bliss, Y. Barniv, "Real-time implementation of a dynamic programming algorithm for multi-target tracking", Saxpy Computer Corporation Final Report, no. pc1009fr.tex, US Army SDC contract DASG-60-87-C-0062, 61 pages, January 1988.

[Wilson88] R. Wilson, H. Knutsson, "Uncertainty and Inference in the Visual System", *IEEE Trans. SMC*, vol. 18, no. 2, pp. 305-312, 1988.

[Wolfe85] W. L. Wolfe, G.J. Zissis, The Infrared Handbook, revised edition, Office of Naval Research, Department of the Navy, 1985.

[Yarbus67] Alfred L. Yarbus, Eye Movements and Vision, Plenum Press, 1967.

[Yeshu89] Y. Yeshurun, E.L. Schwartz, "Shape description with a space-variant sensor: algorithms for scan-path, and convergence over multiple scans", *IEEE Trans. PAMI*, vol. 11, no. 11, pp. 1217-1222, November 1989.

[Zeevi88] Y. Y. Zeevi, N. Peterfreund, E. Shlomot, "Pyramidal Image Representation in Nonuniform Systems", *SPIE*, vol. 1001, *Visual Communications and Image Processing*, pp. 563-571, 1988.

[Zimme86] H.-G. Zimmer, "Vectorial features in pyramidal image processing", in Pyramidal Systems for Computer Vision, V. Cantoni, S. Levialdi (eds.), Springer-Verlag, pp. 299-310, 1986.

[Zucker84] S. W. Zucker, P. Parent, "Multiple-size operators and optimal curve finding", in Multiresolution Image Processing and Analysis, A. Rosenfeld (ed.), Springer-Verlag, pp. 6-35, 1984.